Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach With Robust Feature Selection

PRERNA CHIKERSAL, Carnegie Mellon University, USA
AFSANEH DORYAB, University of Virginia, USA
MICHAEL TUMMINIA, University of Pittsburgh, USA
DANIELLA K. VILLALBA, Carnegie Mellon University, USA
JANINE M. DUTCHER, Carnegie Mellon University, USA
XINWEN LIU, Carnegie Mellon University, USA
SHELDON COHEN, Carnegie Mellon University, USA
KASEY G. CRESWELL, Carnegie Mellon University, USA
JENNIFER MANKOFF, University of Washington, USA
J. DAVID CRESWELL, Carnegie Mellon University, USA
MAYANK GOEL, Carnegie Mellon University, USA
ANIND K. DEY, University of Washington, USA

We present a machine learning approach that uses data from smartphones and fitness trackers of 138 college students to identify students that experienced depressive symptoms at the end of the semester and students whose depressive symptoms worsened over the semester. Our novel approach is a feature extraction technique that allows us to select meaningful features indicative of depressive symptoms from longitudinal data. It allows us to detect the presence of post-semester depressive symptoms with an accuracy of 85.7% and change in symptom severity with an accuracy of 85.4%. It also predicts these outcomes with an accuracy of >80%, 11-15 weeks before the end of the semester, allowing ample time for preemptive interventions. Our work has significant implications for the detection of health outcomes using longitudinal behavioral data and limited ground truth. By detecting change and predicting symptoms several weeks before their onset, our work also has implications for preventing depression.

CCS Concepts: • **Human-centered computing → Human computer interaction**; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: Mobile Sensing, Mobile Health, Mental Health, Depression, Machine Learning, Feature Selection

Authors' addresses: Prerna Chikersal, Carnegie Mellon University, Pittsburgh, PA, USA, prerna@cmu.edu; Afsaneh Doryab, University of Virginia, Charlottesville, VA, USA, ad4ks@virginia.edu; Michael Tumminia, University of Pittsburgh, Pittsburgh, PA, USA, mjtumminia@pitt.edu; Daniella K. Villalba, Carnegie Mellon University, Pittsburgh, PA, USA, dvillalb@andrew.cmu.edu; Janine M. Dutcher, Carnegie Mellon University, Pittsburgh, PA, USA, janine@cmu.edu; Xinwen Liu, Carnegie Mellon University, Pittsburgh, PA, USA, xinwenl@andrew.cmu.edu; Sheldon Cohen, Carnegie Mellon University, Pittsburgh, PA, USA, scohen@cmu.edu; Kasey G. Creswell, Carnegie Mellon University, Pittsburgh, PA, USA, kasey@andrew.cmu.edu; Jennifer Mankoff, University of Washington, Seattle, WA, USA, jmankoff@cs.washington.edu; J. David Creswell, Carnegie Mellon University, Pittsburgh, PA, USA, creswell@cmu.edu; Mayank Goel, Carnegie Mellon University, Pittsburgh, PA, USA, mayankgoel@cmu.edu; Anind K. Dey, University of Washington, Seattle, WA, USA, anind@uw.edu.

XXXX-XXXX/2020/9-ART $15.00
https://doi.org/10.1145/1122445.1122456

ACM Reference Format:
Prerna Chikersal, Afsaneh Doryab, Michael Tumminia, Daniella K. Villalba, Janine M. Dutcher, Xinwen Liu, Sheldon Cohen, Kasey G. Creswell, Jennifer Mankoff, J. David Creswell, Mayank Goel, and Anind K. Dey. 2020. Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach With Robust Feature Selection. 1, 1 (September 2020), 44 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Depression is a common and serious mental health disorder that is especially prevalent among college students. In 2013, the percentage of college students in the United States that reported having difficulty functioning in the last 12 months due to depression was over 33%[1]. Depression has been found to affect academic participation, productivity, and performance [35, 37], and may double the likelihood of dropping out from college [30]. Further, depression is the most common disorder among people with suicidal behaviors [40, 41, 52]. It is estimated that approximately 11.2% of undergraduates seriously considered suicide and 2.1% attempted suicide in 2015-2016[2].

Although treatment for depression is effective and includes a variety of methods, such as psychotherapy and medication, a large number of affected students do not seek treatment [29, 34]. Commonly reported barriers to seeking treatment include the belief that stress is a normal part of student life and treatment is not needed. Furthermore, students may not be aware that they are experiencing not only stress, but also depression [17]. Tools used to monitor the severity of depressive symptoms rely on periodic self-reports that are subjective and if administered too often may reduce compliance. Hence, there is a need to develop more efficient methods to monitor and identify changes in depressive symptoms in college students, and predict future depressive episodes.

Built-in sensors on mobile phones and wearable fitness trackers allow us to passively and unobtrusively collect information such as location, communication, environment, phone usage, physical activity, and sleep. Previous work has shown that such information is linked to depressive symptoms, such as social isolation and sleep disturbances [4]. Measuring the severity of depressive symptoms using such sensors could enable continuous depression detection, prediction before onset, and longitudinal symptom monitoring in-the-wild. Ultimately, it creates the potential for technology-mediated real-time interventions that support the diagnosis, treatment, and prevention of depression. As a result, over the past few years, researchers have conducted several studies that use statistics to understand the relationship between sensor data from phones and wearables, and depression [9, 23, 39, 59, 60, 74]. A growing body of research also focuses on using machine learning to detect depression using sensor data [11, 32, 60, 73, 77], and there has been some initial work on predicting depression in advance as well [11].

Depression, however, is a long-term health problem that needs to be continuously monitored and managed. Although mobile and wearable technology make the long-term monitoring of depression possible, some issues remain. Machine learning (ML) methods used for detecting and predicting depression rely on subjective ground truth acquired through psychological questionnaires such as BDI-II [67]. ML models are trained to detect these scores and their output is compared with these scores to measure their accuracy of prediction. Obtaining ground truth from users with depression or any mental health problem frequently over a long period of time is not sustainable as frequent requests to complete questionnaires will over time become an extra burden especially when the user is experiencing severe symptoms. Nevertheless, so far, all existing research in detecting and inferring depression has relied on frequent measurement of depression status (*e.g.*, every week). Further, while existing research has evaluated ML methods for detecting the presence of

[1] http://www.apa.org/monitor/2014/09/cover-pressure.aspx
[2] http://www.acha-ncha.org/reports_ACHA-NCHAIIc.html

depressive symptoms, whether or not these methods can capture changes in depressive symptoms is unexplored.

In this paper, we present a machine learning approach that uses data from mobile and wearable sensors to detect and monitor depression and change in depression at any time point, with limited ground truth data. Although our approach can be generalized to any chronic and longitudinal health problem, we evaluate it in the context of depression. We use data from smartphones and wearable fitness trackers from 138 students at an R-1 Carnegie-classified US University to identify students who experienced depressive symptoms or whose depressive symptoms worsened by the end of a semester. Our machine learning approach advances the research in mobile health and analysis as follows:

(1) To build machine learning models that can make accurate predictions from long-term data without frequent ground truth acquisition (in our case only two measurements at the beginning and end of semester), data needs to be processed and aggregated without losing key behavioral information during different time periods that may be useful in detecting and predicting depression. Therefore, we extract fine-grained features to capture behavioral markers in different time windows with varying granularity during the day, week, and semester. Although this step results in a number of features (>60,000) that is significantly larger than the number of samples (138 students), the hierarchical and incremental modeling component and stable feature selection in the pipeline are capable of identifying the most significant features, *i.e.*, features that are commonly chosen in most validation runs. We evaluate our approach by identifying students that have post-semester depressive symptoms using data collected over one semester (16 weeks) from the smartphones and fitness trackers of 138 college students, and achieve an accuracy of 85.7%. We demonstrate that our method outperforms off-the-shelf ML methods such as Lasso and K-Nearest Neighbors.
(2) We also evaluate our approach on its ability to detect change in depressive symptoms. To the best of our knowledge, our work is the first to detect change in depressive symptom severity without any knowledge of the students' initial or previous depression severity. We detect whether students' depressive symptom severity changed with an accuracy of 85.4%.
(3) Previous work on prediction has only looked into predicting depression 0-2 weeks in advance and it may not leave enough time for interventions [11]. Our work is the first to demonstrate that it is possible to predict depression several weeks in advance. We are able to identify students who will have depressive symptoms by the end of the semester with an accuracy of 81.3%, 11 weeks before the semester ends.

2 RELATED WORK

The Diagnostic and Statistical Manual of mental disorders (DSM-5) [4] describes several depressive disorders, most prevalent of which are Major Depressive Disorder (MDD) and Persistent Depressive Disorder (PDD). People with these disorders experience similar symptoms over different periods of time (*e.g.*, at least 2 weeks for MDD and at least 2 years for PDD). These symptoms include *depressed mood, diminished interest or pleasure in almost all activities, sleep disturbances* such as insomnia or hypersomnia, *psychomotor agitation or retardation, fatigue or loss of energy, feelings of worthlessness or guilt, diminished concentration,* and *recurrent thoughts of suicide*. Many of these symptoms manifest as verbal, non-verbal or daily behaviors [22] and can be passively sensed with limited user involvement.

Automated techniques for identifying depressive symptoms can be grouped based on the type of behavioral symptoms they sense – verbal [18, 58], non-verbal [1–3, 15, 38, 62–64, 68], or daily. Our approach focuses on daily behaviors that can be sensed using smartphones and fitness trackers, which allows for depression detection and longitudinal symptom monitoring.

Daily behaviors are related to communication, movement patterns, smartphone use, sleep, and physical activity, which can be sensed using sensors embedded in smartphones and fitness trackers. Features indicative of daily behaviors can be extracted from sensor data to capture behavioral symptoms of depression. Previous research on this topic either uses statistical analysis to explore the relationship between these features and depression or uses these features to build machine learning models to detect depression.

2.1 Exploring the statistical relationship between behavioral features and depression

Doryab *et al.* [23] explored detection of behavior change in people with major depression from smartphone data. Their pilot study of three participants (2 female and 1 male) over 4 months found an inverse relationship between the number of outgoing calls and depression scores over time with the male patient, and a direct relationship between the number and duration of outgoing calls and depression scores over time with the female patients. A study with 216 college students [39] demonstrated a direct relationship between Internet use and depression, *i.e.*, students with depressive symptoms used the Internet significantly more than non-depressed students. They

Table 1. Related Work for Depression Detection. For this paper (last row), note that "all" results are obtained using all features, while "best" results are obtained via a feature ablation study (see section 4).

Reference	**Part.**	**Duration**	**Sensors**	**Outcome**	**Accuracy**	**Other Metrics**
[60]	28 adults	2 wks	Location (only 1 feat.)	Dep. at end of 2 wks	86.5%	
[11]	28 adults	avg. 10 wks/user	Location	Detecting dep. over different periods of time, and predicting dep. 1-14 days in advance.		Sensitivity= 0.71/Specificity= 0.87
[73]	36 people	Variable	Smartphone sensors	Dep. biweekly	61.5%	F1=0.62
[32]	79 col. age	7-8 mos	Location	Clinical dep. biweekly		F1=0.82
[77]	68 col. studs	18 wks	Smartphone sensors (light, GPS, accelerometer, microphone, screen status) & heart rate sensor	Dep. weekly		F1=0.75
[45]	28 adults	avg. 10 wks/user	Location	Detecting dep. over different periods of time. No early prediction.		Sensitivity= 0.77/Specificity= 0.91
This paper	**138 col. studs**	**16 wks**	**Smartphone sensors (bluetooth, calls, GPS, microphone, screen status) & wear. fitness tracker (steps, sleep**	Post-semester dep.	85.7% (best); 82.3% (all)	F1=0.82 (best); 0.78 (all)
				Change in dep.	85.4% (best); 75.9% (all)	F1=0.80 (best); 0.67 (all)
				Explored predicting the above 2 outcomes 1-15 weeks in advance. Results: > 80% accuracy 11-15 weeks ahead of the end of semester, for both prediction problems.		

also switched more frequently between email, chat rooms, social media, video watching, and games. Saeb *et al.* [60] explored the relationship between depression severity score and mobile data including location traces and phone usage in 28 adults over a two-week period and found significant correlations between participants' depression scores (from a standardized assessment) and Location features such as location variation, regularity in movement over days ("circadian movement") and evenness in time spent across locations ("location entropy"). They also found significant correlations between phone usage features such as usage duration and frequency. They replicated the same results using Location features on another dataset [74] containing data from 48 college students over a 10-week period [59]. This dataset was originally collected as part of the StudentLife study at Dartmouth [74] which revealed significant correlations between depression scores and sleep duration, conversation duration, as well as frequency and number of collocations. Further analysis of the dataset showed significant relationships between change in depression scores and features such as sleep duration, speech duration, and geospatial activity (from locations and WiFi scans) [9].

2.2 Detecting depression

The statistical relationships described above suggest that machine learning models could be used to detect depression. As summarized in Table 1, existing work has made important strides in this domain. Saeb *et al.* [60] were able to achieve a leave-one-participant-out accuracy of 86.5% for distinguishing between participants with depressive symptoms and those without depressive symptoms. However, they collected data from 28 adults over a short two-week period and used only one feature from the Location sensor in their machine learning model. Further, cross-validation was not used for feature selection, thus reducing the generalizability of their model. Canzian and Musolesi [11] trained personal models for each of their 28 users using features related to mobility patterns from location data, to detect periods in which users experience an unusual depressed mood. Their models achieved high sensitivity and specificity values, which means that for most of the users, they were able to detect periods of depressed mood (related to sensitivity) while generating few false alarms (related to specificity). They also extended their approach to predicting depressive symptom severity 1-14 days in advance. Wahle *et al.* [73] detect biweekly depression in 36 participants over 2-10 weeks using a very limited set of features from location, physical activity, phone usage, calls, texts, and WiFi scans, and achieve an accuracy of 61.5%. Farhan *et al.* [32] detect biweekly depression in 79 college-age participants over 7-8 months using location data as input to their model and clinical evaluations as their ground truth, and achieve a F1 score of 0.82. Wang *et al.* [77] detect depression on a week-by-week basis using features from smartphone and wearable data as input and weekly subjective assessments as ground truth from 68 undergraduates over two 9-week terms, and achieved 81.5% recall and 69.1% precision. In addition to some of the above features, they used campus-specific features such as time spent in dorm and time spent at study places. Mehrotra and Musolesi [45] used autoencoders for automatically extracting features from the raw GPS data, and achieved better results than "hand-crafted" location features.

All of this earlier work has heavily relied on frequent assessment of depression (weekly or biweekly). As mentioned previously, in real world situations, the mental health status of individual people is often unknown which makes the above mentioned approaches less usable and realistic. In this paper, we address this specific issue through developing a machine learning pipeline capable of detecting depression without frequent ground truth data. Further, while subjective measures for depression are evaluated for their "sensitivity to change" [8, 42, 50], the same has not been done for depression models based on passive sensing. That is, we do not know if existing ML methods for depression detection work well because they capture transient depressive symptoms or latent characteristics known to increase the risk of depression (*e.g.* early major life events [49], thought patterns [71]). In this paper, in addition to detecting post-semester depression, we detect change in depression, thereby resulting in ML models that capture transient depressive symptoms.

Finally, predicting depression in advance is a very useful task as it can allow us to intervene *before* the onset or worsening of symptoms. Subjective measures for depression are designed to measure symptom severity at a particular time by directly asking the participant about their symptoms. Passive sensing models, however, have the potential to do more than that, as they may be able to capture early behavioral signs of depression that even the participant may not be aware of. Other than the study in [11] which attempted to predict depression 0-2 weeks in advance, we are unaware of any research in early prediction of depression. With our approach, we can predict the post-semester depression with an accuracy of > 80% as early as week 5 of the 16 weeks-long semester, giving clinicians a larger window of time for interventions.

In the following sections, we describe our approach in detail, starting with data collection.

3 DATA COLLECTION

In this section, we describe the participant recruitment and the data collection process (including participant-reported depression measures and passively sensed data from smartphones and fitness trackers).

3.1 Participants and Recruitment

Participants in the study were from a pool of first-year undergraduate students at a Carnegie-classified R-1 University in the United States. Students were eligible to participate in the study if they were enrolled as a full-time student on campus for the semester and owned a data plan-enabled smartphone running iOS or Android. The research team advertised the study *via* emails and posts to student mailing lists and Facebook groups. Students were invited to our lab to be screened for eligibility, provide informed consent, download a mobile application to track sensor data from their smartphones and receive a Fitbit Flex 2 to track steps and sleep. After enrollment, the students completed an initial depression questionnaire online. They also gave us the phone numbers of their top-10 family members, friends on-campus, and friends off-campus, which were used to compute certain "calls"-related features (see section 4.1.2). Data was collected from smartphone and Fitbit sensors as described in Section 3.3 and was continuously recorded over the duration of the study: one semester (16 weeks).

Out of the 188 first-year college students initially recruited, 138 completed the study and the depression questionnaires at the beginning and the end of the study. The questionnaires were delivered via email and administered using Qualtrics – an online survey platform [56]. For their participation, the participants were allowed to keep the Fitbit Flex 2 and were compensated up to USD $205 spread over different points in time – $10 after the baseline appointment, $20 after the pre-semester depression questionnaire, $25 after week 1, $40 after week 7, $60 after week 15, $25 after post-semester depression questionnaire, and $25 bonus for compliance.

3.2 Participant-reported Depression Measures (Ground Truth)

The Beck Depression Inventory-II (BDI-II) [8, 25] is a widely used psychometric test for *measuring the severity of depressive symptoms*, and has been validated for college students [25, 67]. It contains 21 questions, with each answer being scored on a scale of 0-3. *Higher scores indicate more severe depressive symptoms.* For college students, the cut-offs on this scale are 0-13 (no or minimal depression), 14-19 (mild depression), 20-28 (moderate depression) and 29-63 (severe depression) [25].

The semester spanned over 16 weeks towards the end of which exams start and continue into the 17$^{\text{th}}$ week. Since we expected compliance for answering the post-semester depression questionnaire during exams to be low, we concluded the study at the end of week 16. Participants answered questions from BDI-II at the beginning (week 1) and at the end (week 16) of the semester, which gave us their pre-semester and post-semester depression scores indicating the severity of depressive symptoms. From these scores, we calculated ground truth for two outcomes, as follows:

(1) *Post-semester Depression (Binary):* All participants with no or minimal depression (post-semester BDI-II score < 14) at the end of the semester were classified as *"not having depression"*. While all participants with mild, moderate, or severe depression (post-semester BDI-II score >= 14) at the end of the semester were classified as *"having depression"*.
(2) *Change in Depression (Binary):* We compare the pre-semester depression severity levels to the post-semester depression severity levels to obtain the *change in depression severity levels*. Using the standardized thresholds listed above, we assessed both pre-semester and post-semester BDI-II scores as being in one of four levels: no or minimal, mild, moderate, or severe. The depression severity levels did not improve for any participant. If there was no change in depression severity levels for a participant, the participant's *"Change in Depression"* was classified as *"did not worsen"*, otherwise it was classified as *"worsens"*.

3.3 Passive Data Collection

We installed the AWARE framework [33] – a data collection mobile application with supporting backend and network infrastructure to collect sensor data unobtrusively from students' smartphones. This enabled us to record nearby bluetooth addresses, location, phone usage (*i.e.*, when the screen status changed to *on* or *off* and *locked* or *unlocked*), and call logs for incoming, outgoing and missed calls. In order to assess calls to close contacts, we asked participants to provide phone numbers of family members, friends on-campus, and friends off-campus that they most frequently contact. We also used a conversation plugin for AWARE (same as the one used by [74]) which makes audio inferences such as silence, voice, noise, or unknown. Further, we equipped participants with a Fitbit Flex 2 which records the number of steps and sleep status (asleep, awake, restless, or unknown). Calls, conversation, and phone usage are event-based sensor streams, whereas Bluetooth, location, sleep, and steps are sampled time series. These time series data streams were sampled at different rates, due to the capabilities of the hardware being used. Bluetooth and Location coordinates are sampled at 1 sample per 10 minutes, sleep at 1 sample per minute, and steps at 1 sample per 5 minutes.

Data from AWARE was deidentified and automatically transferred over WiFi to our backend server on a regular basis, and data from the wearable Fitbit was retrieved using the Fitbit API at the end of the study. Participants were asked to keep their phone and Fitbit charged and carry/wear them at all times.

To maintain the participants' privacy and confidentiality, we stored all identifiable information (*e.g.* names, contact information) separate from their deidentified survey and sensor data. Only a few authorized members of the research team had access to the participants' identifiable information. All data sources – identifiable or not were password protected for security. At the University where this research was conducted, the Institutional Review Board (IRB) reviewed, oversaw and approved all procedures.

4 DATA PROCESSING AND ANALYSIS

This section describes the data processing and analysis pipeline, that consisted of 4 main steps:

(1) Feature extraction to acquire sets of behavioral and behavioral change features from different sensors over different time slices (Section 4.1).
(2) Handling missing features (Section 4.2).

(3) Machine learning to detect Post-semester Depression and Change in Depression (Section 4.3), which involved:
  (a) Detecting depression outcomes using 1-feature set models (*i.e.*, models containing features from one sensor).
  (b) Combining detection probabilities given by these 1-feature set models to obtain a final detection label for our two outcomes.
(4) Further, we slightly modified step (3) for different outcomes, different sensor combinations, and to predict future depressive episodes (also in Section 4.3).

This pipeline is illustrated in Figure 1 and explained in the subsections below.

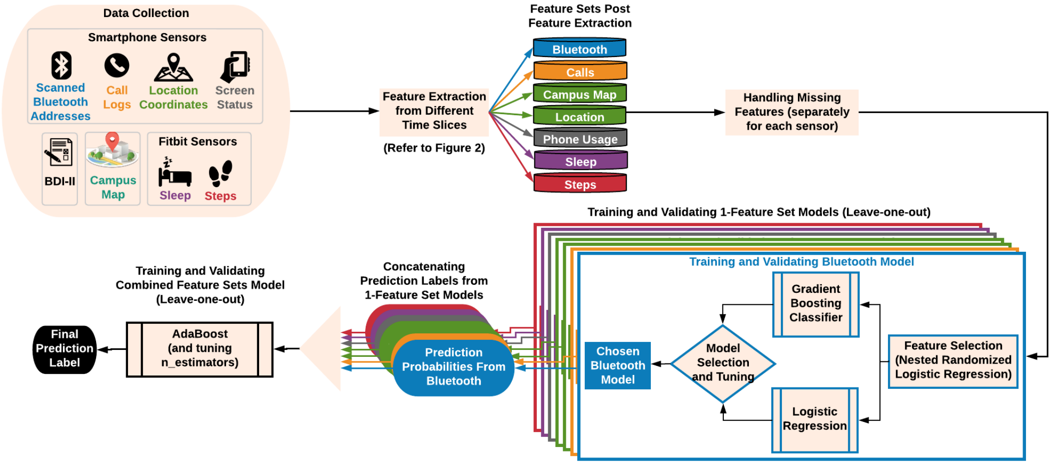

Data Collection
Smartphone Sensors
Scanned Bluetooth Addresses
Call Logs
Location Coordinates
Screen Status
BDI-II
Campus Map
Fitbit Sensors
Sleep
Steps
Feature Extraction from Different Time Slices
(Refer to Figure 2)
Feature Sets Post Feature Extraction
Bluetooth
Calls
Campus Map
Location
Phone Usage
Sleep
Steps
Handling Missing Features (separately for each sensor)
Training and Validating 1-Feature Set Models (Leave-one-out)
Training and Validating Bluetooth Model
Gradient Boosting Classifier
Model Selection and Tuning
Logistic Regression
Feature Selection (Nested Randomized Logistic Regression)
Chosen Bluetooth Model
Prediction Probabilities From Bluetooth
Concatenating Prediction Labels from 1-Feature Set Models
Training and Validating Combined Feature Sets Model (Leave-one-out)
AdaBoost (and tuning n_estimators)
Final Prediction Label

Fig. 1. Pipeline for the data processing and analysis.

4.1 Feature Extraction

We computed *seven* feature sets from the collected data: Bluetooth, Calls, Location, Campus Map, Phone Usage, Steps, and Sleep. These feature sets were chosen because they have the potential to capture depressive symptoms described in the DSM-5 [4]. Location and Campus Map features capture users' mobility patterns; Calls features capture communication patterns; Bluetooth features can reflect both mobility and communication patterns; and Steps capture physical activity. Together they can be strong indicators of social withdrawal and diminished interest or pleasure in almost all activities, especially social and occupational activities. Further, fatigue or loss of energy can cause users to take longer to perform certain tasks, which may also be represented by these features. Sleep disturbances such as insomnia or hypersomnia are often present in people with depression [53]. Depression also causes diminished concentration which can affect phone usage [20, 43].

Due to some technical issues with AWARE's conversations plugin, many conversation inferences were missing. Hence, we used available conversation inferences to inform a Campus Map feature called social duration (explained later), but did not extract a Conversations feature set.

The feature extraction approach for each of the seven feature sets is described in Sections 4.1.1–4.1.7. These features were extracted over different temporal slices (see section 4.1.8). We also extracted behavioral change counterparts for every feature (see section 4.1.9). As a result, we obtained 14 feature matrices for the seven feature sets and their behavioral change counterparts as explained in section 4.1.10. *We did not include pre-semester depression scores or labels as features in any of our models.*

4.1.1 *Bluetooth Feature Set.*

Bluetooth features were calculated from the scanned bluetooth addresses recorded by the Bluetooth sensor in the smartphone, and can be used to sense the user's social context [12, 13, 27, 51, 79]. While a relationship between a bluetooth feature and depression has been found [74], bluetooth features have not been used to *detect* depression.

Scanned bluetooth addresses can be clustered into the participant's own devices ("self" – scanned most often), family/ roommate/ office mate's devices ("related" – scanned less often than "self" but more often than "others"), and other people's devices ("others" – scanned least often) to help us estimate how many different people the participant meets, thereby capturing social activity and collocated communication. Since a participant may or may not be living with family or a roommate or be sharing an office, we clustered scanned addresses twice. First, the addresses were clustered into two categories for "self" *vs.* "others" ($K = 2$ clusters), then into three clusters - "self" *vs.* "related" *vs.* "others" ($K = 3$ clusters), and then chose the model which fit the data best out of the two sets. This process is described below.

(1) We calculated the number of days each unique bluetooth address was scanned at least once. That is, $number\_of\_days_{bti}$.
(2) We calculated the average frequency of each unique bluetooth address. That is, $average\_frequency_{bti} = \frac{total\_count_{bti}}{number\_of\_days_{bti}}$.
(3) We Z-normalized the $number\_of\_days_{bti}$ and $average\_frequency_{bti}$ in order to give equal weight to both while optimizing score in step 4.
(4) For each bluetooth address, we computed $Score = number\_of\_days_{bti}+average\_frequency_{bti}$.
(5) We used K-means clustering to cluster *Score* from step 4 for all bluetooth addresses using K=2 and K=3.
(6) We chose the model with K=2 if sum of squared distances between clustered points and cluster centers was smaller than what we get with K=3. Otherwise we chose model with K=3.
(7) If model with K=2 was chosen, the cluster with higher scores contained the participant's own devices ("self"), while the other cluster contained other people's devices ("others"). If the model with K=3 was chosen, the cluster with highest scores contained the participant's own devices ("self"), the cluster with lowest scores contained other people's devices ("others"), and the remaining cluster contained devices of the participant's partners, roommates, or officemates ("related").

Once the bluetooth addresses scanned were clustered into "self" *vs.* "others" or "self" *vs.* "related" *vs.* "others", we extracted the *number of unique devices*, *number of scans of most and least frequent device*, and *sum, average, and standard deviation of the number of scans of all devices* from all devices (*i.e.*, ignoring clusters), "self" devices, "related" devices, or "others" devices.

It is important to note that we do not have the bluetooth addresses of devices belonging to the user or people related to the user. We are using the frequency of occurrence of the devices scanned to heuristically 'guess' these clusters/ categories. Wang *et. al.* [74] used the number of collocated bluetooth devices to estimate the user's social context, however these devices may or may not belong to other people. However, these devices would also include the user's own devices, and hence may not accurately represent the user's social context. By using the frequency of occurrence of these devices to obtain 3 clusters, we build on previous work by attempting to separate the devices that are more likely to (1) belong to the user ("self"), (2) belong to people the user meets/ sees regularly ("related"), and (3) belong to other people ("others"). If the user does not meet many people regularly, then K=2 may fit the data better than K=3, thus giving us devices that are more likely to (1) belong to the user ("self") and (2) belong to other people ("others").

4.1.2 Calls Feature Set. Calls features were calculated using the call logs from the smartphone. We extracted the following features:

Number and duration of incoming, outgoing, and missed calls to everyone, family members, friends off-campus, and friends on-campus, *number of correspondents overall*, and *number of correspondents who are family members, friends off-campus or friends on-campus*.

4.1.3 Location Feature Set. Location features are derived from the Location 'virtual' sensor of the smartphone which uses proprietary algorithms to come up with the best estimate of location based on available GPS, WiFi and Celltower signals. We extracted the following Location features:

Location variance (sum of the variance in latitude and longitude coordinates), *log of location variance*, *total distance traveled*, *average speed*, and *variance in speed*. *Circadian movement* [60] was calculated using the Lomb-Scargle method [55]. It encodes the extent to which a person's location patterns follow a 24-hour circadian cycle. Then, we carried out the following processing steps:

(a) Speed of the person was calculated from the distance covered and time elapsed between two samples. Samples with speed > 1 km/h were labeled as "moving", else "static" [59, 60].
(b) Samples labeled as "static" were clustered using DBSCAN [31] to find significant places visited by the participant. When we clustered all data and extracted each feature per week or per half-semester, we used global clusters. When we first split the data into weeks or half-semesters and then extracted features from each temporal slice, we used local clusters. Temporal slicing is discussed in section 4.1.8.

These steps allowed us to extract: *number of significant places*, *number of transitions between places*, *radius of gyration* [11], *time spent at top-3 (most frequented) local and global clusters*, *percentage of time spent moving*, and *percentage of time spent in insignificant or rarely visited locations* (labeled as -1 by DBSCAN). We also calculated *statistics related to length of stay at clusters* such as maximum, minimum, average, and standard deviation of length of stay at local and global clusters. *Location entropy* and *normalized location entropy* across local and global clusters were also calculated (implemented using the method in [60]). Location entropy will be higher when time is spent evenly across significant places. Calculating features for both local and global clusters allowed us to capture different behaviors related to the user's overall location patterns (global) and the user's location patterns within a time slice (local). For example, time spent at top-3 global and local clusters captures the time spent at places of overall significance to the user and places significant to the user in a particular time slice (*e.g.,* mornings on weekends).

We assume the place most visited by the participant at night to be their home location. To approximate the home location, we performed steps (a) and (b) above on the location coordinates from all nights (12am to 6am) and assumed the center of the most frequented cluster to be the participant's home location center. Since we don't know the radius of the home, we calculated two home-related features *time spent at home assuming home to be within 10 meters of the home location center*, and *time spent at home assuming home to be within 100 meters of the home location center*[3].

4.1.4 Campus Map Feature Set. We also analyze the user's location patterns in relation to their college campus. First, we obtained a campus map of the participants' University. Then, we marked out the campus boundary and different types of buildings on campus by creating polygons on Google Maps using GmapGIS[4]. We annotated six types of buildings and spaces – Greek houses

[3] The 100m threshold is the default geofencing radius used by automation systems like HomeKit and https://www.home-assistant.io/, while the 10m threshold corresponds to the accuracy of GPS in an urban environment [47].
[4] http://www.gmapgis.com/

that hold the most social events, all Greek houses, student apartments, residential halls, athletic facilities, and green spaces. As academic buildings in this University are often collocated with other spaces, we assume any on-campus space not belonging to these six categories to be an academic building. For every location sample, we assigned one of eight *location type labels* (6 building/space types, academic, off-campus). Then, the following features were extracted for each type of space: *time spent at each location type in minutes*, *percentage time spent at each location type*, *number of transitions between different spaces*, *number of bouts (or continuous periods of time) at space*, *number of bouts during which participant spends 10, 20, or 30 minutes at the same space*, and *minimum, maximum, average, and standard deviation of length of bouts at each space.*

Campus map features also include two multimodal features – *study duration* and *social duration*, as implemented by Wang *et al.* [75, 77]. These features fuse data from Location, Phone Usage, Conversation, and Steps sensors. Study duration was calculated by fusing location type labels with data from the phone usage and steps sensors. A participant was assumed to be studying if they spent 30 minutes or more in an academic building while being sedentary (fewer than 10 steps) and having no interaction with their phone. Social duration was calculated by fusing location type labels with data from the conversation sensor. A participant was assumed to be social if they spent 20 minutes or more in any of the residential buildings or green spaces and the conversation sensor inferred human voice or noise for 80% or more of that time. Study duration was only calculated in academic buildings, while social duration was only calculated in residential buildings or green spaces.

4.1.5 Phone Usage Feature Set. Phone Usage features were calculated using the screen status sensor in the smartphone, which recorded screen status (on, off, lock, unlock) over time. We extracted the following phone usage features:

Number of unlocks per minute, *total time spent interacting with the phone*, *total time the screen was unlocked*, *the hour of the days the screen was first unlocked or first turned on*, *the hour of the days the screen was last unlocked, locked, and turned on*, and *the maximum, minimum, average, and standard deviation of length of bouts (or continuous periods of time) during which the participant is interacting with the phone and when the screen is unlocked.* A participant is said to be "interacting" with the phone between when the screen status is "unlock" and when the screen status is "off" or "lock".

4.1.6 Sleep Feature Set. Sleep features were calculated from the sleep inferences (asleep, restless, awake, unknown) over time returned by the Fitbit API[5]. The following features were calculated:

Number of asleep samples, *number of restless samples*, *number of awake samples*, *number of unknown samples* (still detected as sleep), *weak sleep efficiency* (sum of number of asleep and restless samples divided by sum of number of asleep, restless, and awake samples), *strong sleep efficiency* (sum of number of asleep samples divided by sum of number of asleep, restless, and awake samples), *count, sum, average, maximum, and minimum length of bouts during which the participant was asleep, restless, or awake* as well as the *start and end time of longest and shortest bouts during which the participant was asleep, restless, or awake.* We include 3 summary statistics – count, sum, and average length of asleep/ restless/ awake bouts as individual features, despite them being dependent on each other, because we want to consider the "interaction" between these features[6]. For example, say larger average length per asleep bout and smaller number of asleep bouts correlate with better mental health outcomes, the relationship between average length per asleep bout and mental health may still be dependent on the number of asleep bouts. Very high number of asleep bouts could

[5] Sleep captured by Fitbit is accurate +/- 45min [16, 19, 44].

[6] Interaction models are commonly used in statistics (see: http://www.medicine.mcgill.ca/epidemiology/Joseph/courses/EPIB-621/interaction.pdf). For example, let x_1 = mean, x_2 = count. Then, the interaction term is $x_1 x_2$ = sum.

indicate disturbed sleep or polyphasic hypersomnia, such that even with high average length per asleep bout, the mental health outcomes could be poor.

4.1.7 Steps Feature Set. Steps features were calculated from the step counts over time returned by the Fitbit API. The following features were calculated:

Total number of steps and *maximum number of steps taken in any 5 minute period* were extracted as features. Other features were extracted from "bouts", where a "bout" is a continuous period of time during which a certain characteristic is exhibited. Examples of such features include *total number of active or sedentary bouts* [5], and *maximum, minimum, and average length of active or sedentary bouts*. We also calculated *minimum, maximum, and average number of steps over all active bouts*. A bout is said to be sedentary if the user takes less than 10 steps during each 5 minute interval within the bout. As soon as the user takes more than 10 steps[7] in any 5 minute interval, they switch to an active bout.

4.1.8 Temporal Slicing. Our temporal slicing approach helps us extract behavioral features from different time slices. Past work has shown that this approach can better elicit the relationship between a feature and depression. For example, Chow *et al.* [14] found no relationship between depression and time spent at home during 4-hour time windows, but they found that people who are more depressed tend to spend more time at home between 10:00 AM and 6:00 PM. Similarly, Saeb *et al.* [59] found that the same behavioral feature calculated over weekdays and weekends can have a very different effect on depression.

Each feature described in Sections 4.1.1–4.1.7 was extracted from 45 temporal slices or time segments as illustrated in figure 2. First, we fetched all available data (spanning over multiple days of the study) from a certain epoch or time of the day (all day, night *i.e.*, 12am-6am, morning *i.e.*, 6am-12pm, afternoon *i.e.*, 12pm-6pm, evening *i.e.*, 6pm-12am) and for certain days-of-the-week (all days of the week, weekdays only *i.e.*, Monday-Friday, weekends only *i.e.*, Saturday-Sunday). Then, we calculated features from this data aggregated over different levels of granularity (whole semester, two halves of the semester, weekly). Since there are 5 epochs, 3 days-of-the-week segmentations, and 3 levels of granularity, we get $5 \times 3 \times 3 = 45$ time slices. Each location feature is calculated over these 45 time slices. Note that the two halves of the semester are not perfect halves. For simplicity, we refer to weeks 1-6 as the first half (before midterms) and weeks 7-16 (midterms and after midterms) as the second half. We also investigated the effect of removing the spring break weeks (week 8 and 9 as the spring break was mid-week 8 to mid-week 9) on detecting the two outcomes, and while our findings were inconclusive, this may be worthy of future study.

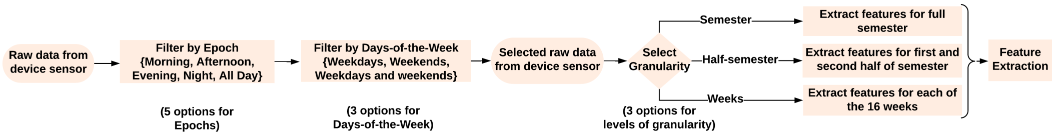

Raw data from device sensor
Filter by Epoch {Morning, Afternoon, Evening, Night, All Day}
(5 options for Epochs)
Filter by Days-of-the-Week {Weekdays, Weekends, Weekdays and weekends}
(3 options for Days-of-the-Week)
Selected raw data from device sensor
Select Granularity
(3 options for levels of granularity)
Semester
Half-semester
Weeks
Extract features for full semester
Extract features for first and second half of semester
Extract features for each of the 16 weeks
Feature Extraction

Fig. 2. For each sensor, each feature was extracted from 45 time slices. First, raw data from the device sensor was preprocessed and then filtered by an epoch and a days-of-the-week option. Features (Let NS be number of features derived from each sensor) were then extracted from the selected raw data according to 3 levels of granularity – per semester (NS features), per half-semester ($2 * NS$ features), and per week ($16 * NS$ features).

[7] A threshold of 10 steps is often used to ignore 'false steps' [61, 69]. Previous work has also used 10 steps as a threshold to detect sedentary behavior [36, 70].

4.1.9 Behavioral Change Features. Behavioral change features capture changes in behaviors over 16 weeks. These features can be abstractly characterized as the change in slope for each behavioral feature over the semester. For this purpose, we only use features computed weekly (*i.e.*, using granularity “weeks”). This gives us 15 time slices (for 5 epochs × 3 days-of-the-week options) for which we have weekly values of every behavioral feature described in Sections 4.1.1–4.1.7. We compute the behavioral change feature for each behavioral feature using their weekly values over 16 weeks. We follow the same method employed by [75], to test whether their approach works on our dataset:

- *Slope:* We fit a linear regression model to the values of the feature over 16 weeks. “Slope” is the slope of this linear regression line.
- *Slope first half and second half:* We fit two separate linear regression models to the values of the feature over weeks 1-6 (*i.e.*, before midterms) and weeks 7-16 (*i.e.*, midterms and after midterms) of the semester. “Slope first half” and “Slope second half” are the slopes of these linear regression lines.
- *Breakpoint:* Each student’s breakpoint is the week after which the student’s behavior (represented by the feature value) begins to change. This is calculated by fitting a piecewise linear regression model with two segments with each of the 16 weeks as a breakpoint. “Breakpoint” is the week that when used as a breakpoint gives the best model as determined by Bayesian Information Criterion (BIC).
- *Slope before and after Breakpoint:* A piecewise linear regression model with two segments is fit to the feature values over 16 weeks with the final “Breakpoint”. The slope of the first line segment is “Slope before Breakpoint” and the slope of the second line segment is the “Slope after Breakpoint”.

4.1.10 Defining the Feature Matrix. After feature extraction we obtain a feature matrix for each of the seven feature sets derived from different sensors, as well as their behavioral change counterparts (*i.e.*, 14 feature matrices in total). In each of these feature matrices, each sample or record contains features extracted from 1 student. We aggregate our features over different timeslices (see section 4.1.8) – over different weeks, in the two ‘halves’ of the semester, and across the whole semester. The features from all these time slices are concatenated to form the feature vector for each student. By investigating features from individual weeks, we aim to capture the variability in a person’s behavior in different time periods. For example, the midterm week may have a greater impact on depression than the Spring break week.

4.2 Handling Missing Features

Missing features are the result of missing data. While we occasionally miss data from all sensors due to non-semantic reasons (*i.e.*, technical issues such as the phone/ app stopped working, the data was not transferred on time, or the server was down, and compliance-related issues such as the user withdrew permissions for the app), we often miss data due to semantic reasons. For example, if the user does not sleep at all during a time period, we will get no sleep data. If the user does not make any calls during a time period, we will get no calls data. Hence, instead of completely ignoring missing data, since we do not know if it was not collected or whether it did not exist to be collected, we have tried to encode it in our features.

A feature (*i.e.*, feature value during a temporal slice) being missing for a large number of people can indicate non-semantic issues such as server-side problems. Hence, we excluded all features that were missing for more than 30 participants. Further, a participant missing a very large number of features can indicate non-semantic issues such as the phone/ app not working, or that they withdrew permissions. Hence if a participant was missing more than 20% of all features from a feature set, we removed that participant. The “30 participants” and ”20% features” thresholds were determined empirically by plotting the number of participants and features remaining for different threshold values and observing where the curve falls off. All the remaining missing features were imputed as “-1” as their “missingness” may be due to semantic reasons and can be useful information for the classifier. The same features calculated over different time slices were viewed independently, such that if a feature was missing for a week for over 30 people, we only removed that feature from that week. In the end, we were left with roughly 79-110 participants and thousands of features for every feature set. The exact numbers were different across feature sets as missing features in each feature set were handled separately. That is, a participant was excluded from a feature set only if they were missing 20% or more of the total number of features in that feature set.

Appendix A analyzes the ‘missingness’ of features.

4.3 Modeling

We use machine learning to build detection models for depression. Our modeling approach consists of the steps below, *each using leave-one-out cross-validation to minimize over-fitting*. That is, we train a separate model to detect an outcome for each participant, and that model does not include the participant in question, during feature selection or training. It is important to remember that each sample contains features from 1 participant only (recall section 4.1.10), such that leave-one-out or leave-one-sample-out is actually leave-one-person-out.

Our model generation process uses the following steps:

(1) *Stable Feature Selection* using Randomized Logistic Regression while leveraging the semantic structure of the temporal slices (section 4.3.1).
(2) *Training and Validating 1-Feature Set Models* for each of the seven feature sets: Bluetooth, Calls, Campus Map, Location, Phone Usage, Sleep, and Steps (section 4.3.2).
(3) *Obtaining the Final Label for the Outcome* by combining detection probabilities from 1-feature set models (section 4.3.3).
(4) *Classifying Different Outcomes* by slightly modifying the pipeline to detect post-semester depression, and change in depression (section 4.3.4).

We describe these steps in the following sections.

4.3.1 Feature Selection.

After handling missing data, we have 79-110 people (depending on the sensor used) and thousands of features for each feature set. So, the sample size is very small in comparison to the number of features. Hence, feature selection is a crucial step of the pipeline. Moreover, it is essential to select stable features, that is the set of selected features should remain stable when we remove or replace a small number of people. For this purpose, we tried a number of feature selection methods[8] but all of them selected unstable features. That is, the features selected greatly varied across cross-validation folds.

Randomized Logistic Regression [46] is a method that creates several random subsamples of the training dataset (200 in our case), computes a logistic regression on each subsample, and selects features by optimizing their importance across all subsamples. That is, a feature is selected if the average of its logistic regression coefficients across all subsamples is above a specified selection threshold, which is treated as a model parameter and tuned during cross-validation. This usually results in a stable set of selected features. However, in our case, since the number of features in each feature set is significantly larger than the sample size, randomized logistic regression also did not work.

[8] We selected features using recursive feature elimination or that give k highest scores from the model, p-values below alpha based on a FPR test, p-values below alpha based on ANOVA test, and p-values below alpha based on Pearson’s correlation.

To address this problem, we decomposed our feature space for each feature set (*e.g.*, for bluetooth) by grouping features from the same time slices, and performed randomized logistic regression on each of these groups. The selected features from all groups (*i.e.*, all time slices) were then concatenated to give a *new and much smaller* set of features. Then, randomized logistic regression was performed again, this time on this *new* set of features to extract the final selected features for the feature set, thereby *nesting* the process. We call this method Nested Randomized Logistic Regression[9], and used it to extract selected features for each of the seven 1-feature set models.

This method was performed in a *leave-one-out manner* such that the model used to detect an outcome for a person did not include that person during the feature selection process.

4.3.2 Training and Validating 1-Feature Set Models (Model Selection and Tuning). For each feature set, we built a model of the selected features from that feature set to detect an outcome. We used leave-one-out cross-validation (same as leave-one-person-out – see section 4.1.10) to choose the model and parameters for that model. We tried two types of learning algorithms – Logistic Regression and Gradient Boosting Classifier. Logistic Regression was tried because our feature selection approach was based on Logistic Regression, while Gradient Boosting was tried because it can perform well on a noisy dataset, learn complex non-linear decision boundaries via boosting and has been effectively used to detect similar outcomes in previous work [76]. We chose the model and model parameters using accuracy as a metric for post-semester and change in depression. The chosen 1-feature set model gave us detection probabilities for each outcome label.

4.3.3 Combining Detection Probabilities from 1-Feature Set Models to Obtain Combined Models. The detection probabilities from all seven 1-feature set models were concatenated into a single feature vector and given as input to an ensemble classifier, *i.e.*, AdaBoost with Gradient Boosting Classifier as a base estimator, which then outputted the final label for the outcome. For post-semester and change in depression, only the detection probabilities of class label "1" were concatenated. The "n_estimators[10]" parameter was tuned during leave-one-out cross-validation to get the best combined model.

We also carried out a *feature ablation study* to analyze the effect that different feature sets have on the performance of the models, thereby understanding their salience. For this purpose, we concatenated detection probabilities from specific 1-feature set models instead of all seven 1-feature set models. We do this for *all possible combinations of 1-feature set models*, in order to analyze the usefulness of each feature set. There are seven 1-feature set models and 120 combinations of feature sets, as total combinations = combinations with 2 feature sets + ... + combinations with 7 feature sets = $\sum_{r=1}^{7} \binom{7}{r} = 120$.

4.3.4 Classifying Different Outcomes. The pipeline described in the sections above was used to detect two outcomes – post-semester depression, and change in depression.

(1) *Post-semester Depression* (Binary – "depression" or "no depression"): We used the pipeline *as described above* without excluding any students and using *accuracy as the metric for model selection and tuning.*
(2) *Change in Depression* (Binary – "depression level did not worsen" or "depression level worsens"): We used the pipeline *as described above* without excluding any students and using *accuracy as the metric for model selection and tuning.*

[9] $Best(F_s) = sel(concatenate[sel(F_{s1}), sel(F_{s2}), \ldots, sel(F_{sT})])$ where F_{si} = features from feature set s and time slice i (e.g. calls features from the mornings on weekdays calculated weekly), T = total number of time slices, and $sel(\ldots)$ is the Randomized Logistic Regression Function. $T = 45$ for regular feature sets and $T = 15$ for behavioral change feature sets. $Best(F_s)$ are the final features selected from feature set s and are given as input to the 1-feature set model for feature set s.
[10] The maximum number of estimators at which boosting is terminated.

4.3.5 Prediction Models for Predicting Future Depressive Symptoms. Being able to predict post-semester depression and change in depression, using data from a limited number of weeks from the beginning of the semester can help us identify students at-risk for depression and get them treatment early. For each week, we trained 1-feature set models on features from the beginning of the semester to the end of that week, and combined all available 1-feature set models to obtain the final outcome label for that week.

To understand this clearly, it is important to recall (from section 4.1.10) that we only have 1 sample per person and the sample or feature vector for each person contains features averaged over different levels of granularity – each week, each half-semester, and the full semester. So when we exclude a week from our analysis, we exclude all features averaged over that week as well as features averaged over the full semester and the half-semester that that week belongs to. For example, in week 1, the feature vector for each person will only contain features averaged over week 1. Whereas, for week 15, the feature vector for each person will contain features averaged over each week from week 1 to 15, as well as features averaged over the first half of the semester. Model parameters were tuned at each time step for all these models.

Canzian and Musolesi [11] investigated the possibility of predicting depression 1-14 days in advance using location features, and achieved acceptable results 13-14 days in advance. In fact, they obtained very similar results at different time points in their analysis. For example, results obtained 13 days in advance were as good as the results obtained 0 days in advance (see Figure 9 of [11]). Based on their results, we hypothesize that we do not need data from 16 weeks to predict depression, and we do not expect the prediction accuracy to monotonically increase as we add features from subsequent weeks. Even though our detection model contains all the features from the previous weeks' prediction models, we hypothesize that it is possible for some prediction models to outperform the detection model since feature selection in machine learning is rarely optimal. Features from certain weeks can add "noise" to the model and reduce the accuracy obtained after those weeks. For example, students may deviate from their regular behavior during weeks 6-9 which include preparing for midterm exams, and spring break, and weeks 15-16 which include submitting final projects and preparing for final exams.

5 RESULTS

In this section, we present our results. First, we report descriptive statistics about the prevalence of depression in our sample of college students. Then, we report the results obtained. *It is important to note that none of our models contained pre-semester depression scores or labels as features.*

5.1 Descriptive Statistics

As mentioned in section 3.2, the four severity levels of depression specified by BDI-II are symptoms reflecting no or minimal depression (score 0-13), mild depression (score 14-19), moderate depression (score 20-28), and severe depression (score 29-63). At the beginning of the semester, 14.5% *i.e.*, 20 out of the 138 participants who completed the study were categorized as having mild (13 participants), moderate (5 participants), or severe (2 participants) depression. At the end of the semester, this number significantly increased to 40.6% *i.e.*, 56 out of the 138 participants were categorized as having mild (25 participants), moderate (19 participants), or severe (12 participants) depression (see figure 3). While the number of students with depression almost tripled by the end of the semester, the post-semester depression rate is comparable to the 33% estimated by the American Psychological Association[11] for US universities. So, depression statistics at the study University are not surprising or unusual.

[11] http://www.apa.org/monitor/2014/09/cover-pressure.aspx

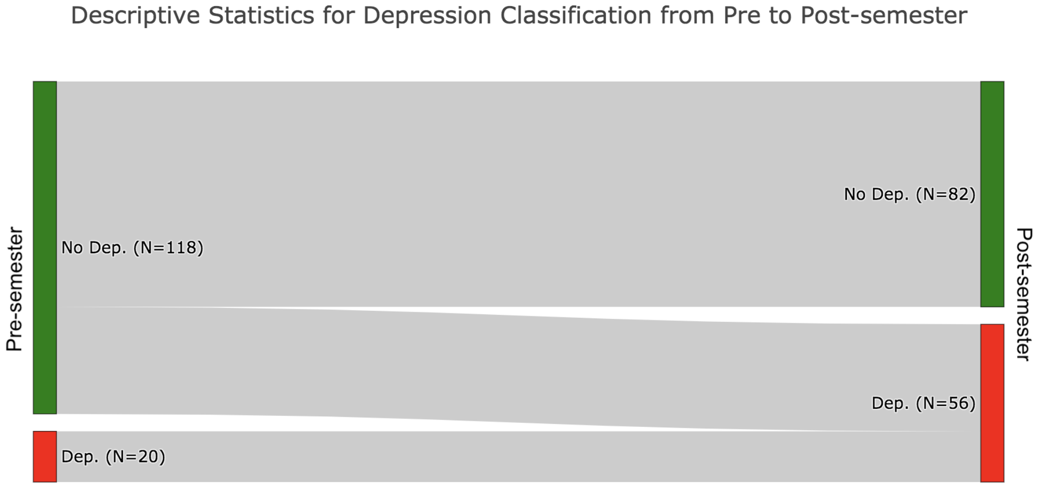


Fig. 3. Shows how depression status ("no dep." vs "dep.") changed from pre to post-semester.

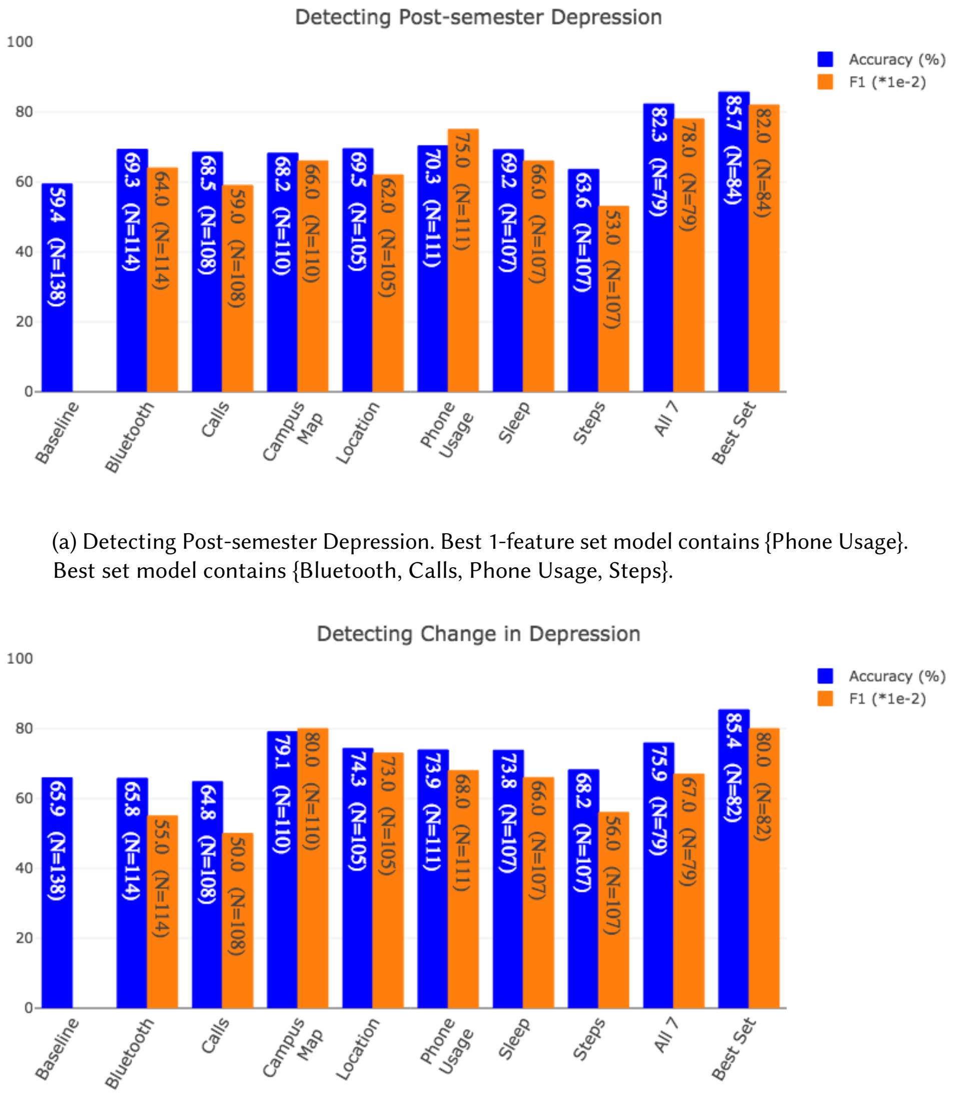
On comparing BDI-II scores from the beginning and end of the semester, we found that the scores of 23 people improved by an average of 2.8, the scores of 99 people got worse by an average of 8.7, and the scores of 16 participants did not change at all. However, on comparing depression severity levels (thresholded scores) from the beginning and the end of the semester, we found that none of the 23 people showed improvement significant enough to improve their depression severity levels. So, the depressive severity levels of none of the participants got better. In fact, depression severity levels did not worsen for 65.9% *i.e.*, 91 out of 138 participants, while they worsened for 34.1% *i.e.*, 47 participants.

5.2 Detecting Post-semester Depression

Figure 4a shows accuracies obtained by the seven 1-feature set models, the 7-feature sets model, and the best set model for detecting *post-semester depression* (*i.e.*, "depression" vs. "no depression"). The *7-feature sets model* is obtained by combining all seven feature sets, while the *best set model* is the model that gives us the best accuracy out of the 120 different combinations of feature sets tried during the feature ablation study. The number of people (*i.e.*, sample size N) may be different for models containing different feature sets since handling missing features for different feature sets will remove a different number of people from the analysis.

If we detect all students as having "no depression" (majority class), we obtain an accuracy of 59.4% (baseline) for detecting post-semester depression. The 7-feature sets model was significantly better than this baseline (using McNemar Test [21, 72], $X^2 = 10.1$ and $p < 0.01$) and obtained accuracy of 82.3% (N = 79). The best accuracy obtained using a 1-feature set model was 70.3% (N = 111) using "Phone Usage". The best set accuracy was 85.7% (N=84) obtained using a model containing 4 feature sets: "Bluetooth", "Calls", "Phone Usage", "Steps" from the feature ablation study (see appendix D). The best set model significantly outperformed the baseline (using McNemar Test [21, 72], $X^2 = 13.4$ and $p < 0.01$), but its performance was not significantly better than the 7-feature sets model (using McNemar Test [21, 72], $X^2 = 0.5$ and $p = 0.48$).

We found that behavioral features are *better* than behavioral change features for detecting post-semester depression. Hence, we only use behavioral features to detect post-semester depression.



(a) Detecting Post-semester Depression. Best 1-feature set model contains {Phone Usage}. Best set model contains {Bluetooth, Calls, Phone Usage, Steps}.

(b) Detecting Change in Depression. Best 1-feature set model contains {Campus Map}. Best set model contains {Bluetooth, Campus Map, Phone Usage, Sleep}.

Fig. 4. Shows accuracies and F1 scores obtained for detecting (a) Post-semester Depression, and (b) Change in Depression. Accuracies and F1 scores are reported for 1-feature set models, the 7-feature set model *i.e.*, model combining detections from all feature sets ("All 7"), and the best set model *i.e.*, the model that gives us the best accuracy during the feature ablation study and thus contains the best set of feature sets ("Best set"). F1 score for (a) is the F1 score of the "depression" class, and F1 score for (b) is the F1 score of the "worsens" class.

The behavioral change features were calculated using the method employed by [75] that assumed that the weekly features have a linear relationship. It is possible that these features don't work well on our dataset because the linearity assumption is false. Therefore, future work should investigate other methods that do not assume linearity for calculating behavioral change features.

5.3 Detecting Change in Depression

Figure 4b shows accuracies obtained by the seven 1-feature set models, the 7-feature sets model, and the best set model for detecting *change in depression* (*i.e.*, "did not worsen" vs. "worsens").

If we detect all students as "did not worsen" (majority class), we obtain an accuracy of 65.9% (baseline) for detecting change in depression. The 7-feature sets model was marginally significantly better than this baseline (using McNemar Test [21, 72], $X^2 = 3.6$ and $p = 0.06$) and obtained an accuracy of 75.9% (N = 79). The best accuracy obtained using a 1-feature set model was 79.1% (N = 110) using "Campus Map". The best set accuracy was 85.4% (N = 82) obtained using a model containing 4 feature sets: "Bluetooth", "Campus Map", "Phone Usage", and "Sleep" from the feature ablation study (see appendix D). The best set model significantly outperformed the baseline (using McNemar Test [21, 72], $X^2 = 12.4$ and $p < 0.01$) and the 7-feature sets model (using McNemar Test [21, 72], $X^2 = 4.5$ and $p < 0.05$).

We found that behavioral features are *better* than behavioral change features for detecting post-semester depression. Hence, we only use behavioral features to detect change in depression.

5.4 Early Prediction of Future Depressive Episodes

This section describes initial results obtained for predicting future depressive episodes using data from the beginning of the semester up to a certain number of weeks until the prediction point. It addresses the question "How early can we predict the two outcomes and with what accuracy?"

Figure 5 contains 2 sub-figures, corresponding to our two outcomes. In each graph **on the left side**, the horizontal axis indicates the week up to which features are included in a model and the vertical axis indicates the accuracy and F1 score that the model obtains. For example, "7" on the horizontal axis means we include features from the start of week 1 to the end of week 7, and the corresponding value on the vertical axis indicates the accuracy a model trained on features from weeks 1 to 7. The best 5 models (with highest accuracies) are labeled. We combine all seven 1-feature set models at each time step, and tune model parameters for them. As mentioned in 4.1.10, we concatenate features from different weeks in order to capture the variability in behaviors across weeks. In the graphs **on the right side**, at each time step, we take the predictions for every participant made by all models up to that time step (as shown in the graph on the left side) and use majority voting to determine the final prediction for every participant. For example, if at least 50% of the models at weeks "1", "2", and "3" predict a participant p as "may have depression", only then will participant p be labeled as "may have depression" in week 3. The graphs on the right side show the final performance obtained when majority voting is applied to the predictions of the models whose performance is shown in the graphs on the left side.

As explained in section 4.3.5, **for the graphs on the left side**, we do not see the prediction accuracy monotonically increase as we add features from subsequent weeks. This is expected and also aligned with previous work [11]. In fact, these prediction models (trained on features from fewer weeks) sometimes outperform the corresponding detection model (trained on features from all weeks) because feature selection in machine learning is rarely optimal. Further, these weeks also have semantic meaning, such that adding data from certain weeks can increase predictive power or introduce noise, thereby affecting accuracy. For example, students have midterms from the beginning of week 7 and 1-2 days into week 8, and spring break during the remainder of week 8 and most of week 9. They typically return to school towards the end of week 9, and weeks 10 and 11 are their first two weeks of regular schoolwork after spring break. While we know what happens in these weeks and our prediction accuracy in the following sections peaks and drops for specific weeks, we cannot associate causality to these results since we do not have any ground truth to support such findings. For example, while most students should have midterms in week 7 or the first 1-2 days of week 8, we don't know the specific days they had their midterms and there may be students who had no midterms at all.

The instability of model performance across weeks makes it harder for the university staff carrying out interventions to trust the output of the model in any one week. Hence, we propose that university staff should look at the predictions from all models previously trained before each time step, and contact participants that are repeatedly labeled as at-risk. Mathematically, this can be achieved using majority voting. In figure 5, **the graphs on the right side** show that after majority voting, performance of the models greatly stabilizes across the 16 weeks. Hence, instead of trusting the output of the prediction model from a specific week, we recommend that the university staff contact at-risk participants every week as long as they have been predicted as at-risk by at least 50% of the models trained until that week.

5.4.1 Predicting Post-semester Depression. The baseline for predicting post-semester depression is 59.4% (see section 4a). Out of the five best prediction models, the model which allows for the earliest prediction needs data from weeks 1 to 5 and achieves an accuracy of 81.3% (N = 80), as shown in figure 5a (left). Hence, we are able to predict post-semester depression with an accuracy significantly better than the baseline as early as the end of week 5. In figure 5a (right), we see that the performance of the prediction models increases quite steadily across the 16 weeks when using majority voting. Therefore, contacting at-risk participants that were labeled as "may have depression" by at least 50% of the models trained until the end of each week, is more reliable and can be repeated every week.

5.4.2 Predicting Change in Depression. The baseline for predicting change in depression is 65.9% (see section 5.3). Out of the five best prediction models, the model which allows for earliest prediction needs data only from weeks 1 to 2 and achieves an accuracy of 88.1% (N = 84), as shown in figure 5b (left). Hence, we are able to predict change in depression with an accuracy significantly better than the baseline as early as the end of week 2. In figure 5b (right), we see that when using majority voting, the performance of the prediction models increases quite steadily across the 16 weeks, with weeks 7 and 9 being the only exceptions[12]. Therefore, contacting at-risk participants that were labeled as "depression may worsen" by at least 50% of the models trained until the end of each week, is more reliable and can be repeated every week.

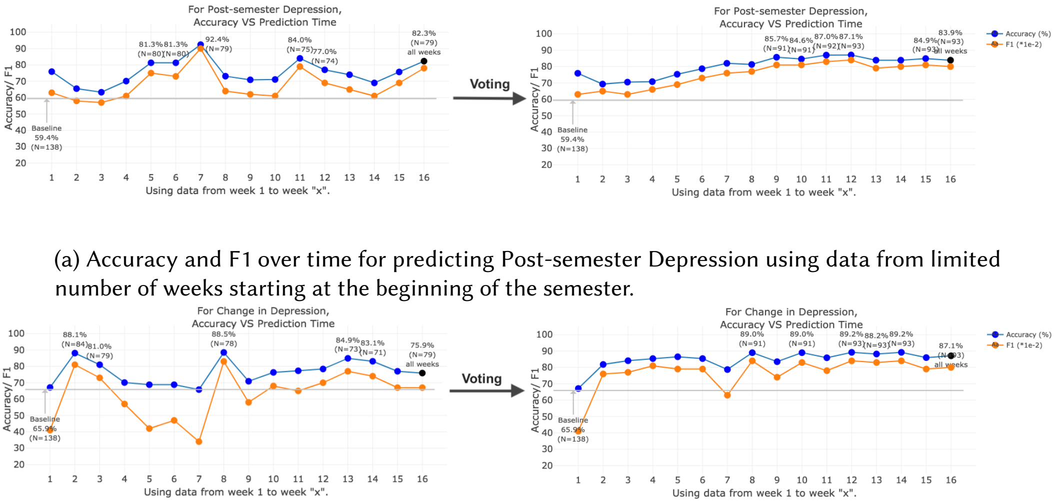
6 DISCUSSION

In this section, we discuss our observations about the selected features, compare our approach with existing ML approaches, and discuss the implications of longitudinal studies, interventions, privacy and technical limitations, and combining daily behaviors with verbal and non-verbal behaviors and genomic data.

6.1 Observations about Selected Features

We made some interesting observations when we informally analyzed the features selected by the best 1-feature set models for detecting each outcome. For this purpose, we took features that were selected in at least one fold, and made different graphs to visualize how many of them came from each feature type, week, epoch, and days-of-the-week. For both outcomes, there is a significant

[12] The drop in performance in weeks 7 and 9 is probably due to atypical behavior during midterms and spring break.



(a) Accuracy and F1 over time for predicting Post-semester Depression using data from limited number of weeks starting at the beginning of the semester.

(b) Accuracy and F1 over time for predicting Change in Depression using data from limited number of weeks starting at the beginning of the semester.

Fig. 5. Accuracies over time obtained when predicting (a) Post-semester Depression, and (b) Change in Depression using data from a limited number of weeks starting at the beginning of the semester (week 1 to week x). In the graphs on the left side, for each time step, detections from all available feature sets are combined to get the corresponding accuracy. In the graphs on the right side, at time step x, majority voting is used on predictions from models at weeks 1 to x (shown on the left) to obtain more stable performance across the 16 weeks.

negative correlation between week number and the number of features selected from each week (post-semester depression: r = -0.94 and p < 0.0001, and change in depression: r = -0.73 and p < 0.0020). That is, more features are selected from earlier weeks and fewer features are selected from later weeks. For all feature sets except sleep and steps, features from nights are selected less often.We interpret this to mean that the participants' social context at night captured using bluetooth, calls, campus map, and location, and their phone usage at night are not predictive of depression. For post-semester depression and change in depression, the most frequently selected features for Bluetooth, calls, and campus map come from afternoons and evenings, and the most frequently selected features for phone usage come from the "all day" epoch. We interpret this to mean that the participants' social context in the afternoons and evenings[13] is the most predictive of depression and change in depression, while their phone usage throughout the day is more predictive of depression and change in depression than their phone usage during specific times of the day. For both outcomes, selected Bluetooth features are related to the devices of "others" and features related to devices of "self" are rarely selected. This shows that the Bluetooth features we calculate to encode proximity to "others" are able to successfully capture depression. Our study is also the first to use Bluetooth features to detect depression. Features such as maximum length of sedentary bouts from steps, maximum length of awake bouts during sleep, time spent in green spaces were most frequently selected for post-semester depression and change in depression. This shows that

[13] Bluetooth, calls, and campus map features from the afternoons and evenings likely reflect social context, as the subjective 'number of social interactions' (in-person and otherwise) reported by participants during weeks 1, 7, and 15, are significantly more in the afternoons and evenings than in the mornings.

long periods of time with no exercise, periods of disturbed sleep at night, as well as time spent outdoors are some of the features that are most predictive of depression and change in depression.

The significant negative correlation between weeks and features selected from each week explains why we are able to achieve an accuracy of >80% very early in the semester using our prediction models, thus enabling depression prevention early in the semester.

Above, we lightly reflected on the features selected, because very little is known about the relationship between behavior *in the wild* and depressive symptoms. Most previous work in this space has only looked at behavior in the therapist's office or behavior self-reported by the participants in retrospect. Hence, validating these features will require qualitatively analyzing participant and clinician experience, which is beyond the scope of this project. That said, the above findings should help explain some of the selected features. To do more would be speculative.

6.2 Comparison with Other Machine Learning Approaches

We compared the results obtained by our novel ML pipeline with two baseline methods – K-Nearest Neighbors and Lasso[14]. For both these methods, we detected Post-semester Depression using models trained on each feature set as well as a model trained on all feature sets. Table 2 shows that our *method outperformed both these methods* for almost every 1-feature set model and for the "all" feature sets model. Comparing the average number of feature selected across all folds, reveals that *our method selected a smaller number of features than the other two methods.* That is, our feature selection approach is more stringent and selects more meaningful features from a large set of features that may often be correlated, as compared to traditional approaches. Our method outperforms traditional approaches for the following reasons:

(1) *Selecting Stable Features by Using Randomized Logistic Regression:* Randomized Logistic Regression selects features by performing Logistic Regression on several subsamples of the training data and selecting features that perform the best across most subsamples. This method leads to a more stable[15] and useful set of selected features as it reduces overfitting by diversifying the training samples. This method and its adaptations hence work well for highly dimensional feature spaces [78, 80].

(2) *Reducing Correlation Between Features During Training by Decomposing the Feature Space Using Data Sources and Temporal Slices:* Some existing ensemble classification methods partition the feature space into smaller subsets using various techniques, learn separate models for each subset, and combine their predictions to get the final prediction. Partitioning the feature space can reduce correlation between features and further diversify the training data that each model is trained on, thereby improving performance [10, 26, 48, 54, 57]. Leveraging the same idea, we decompose the feature space and learn separate 1-feature set models for each data source (*e.g.*, Bluetooth, Location) because we expect different data sources to contain overlapping and correlated behavioral information. For example, step counts (and features derived from step counts) will usually be low when location variance is low. Further, for each 1-feature set model, our novel feature selection approach applies randomized logistic regression on subsets of features from different temporal slices (see section 4.3.1). We do this because features from the same data source can correlate across different temporal slices. For example, a person with low physical activity may have low step count related features in several temporal slices.

[14] Lasso performs regression. We apply use threshold of 0.5 on the score returned by Lasso to achieve binary outcomes.

[15] Features are said to be 'stable' when they don't vary greatly across folds or with minor perturbations of the training data. There is no definite method of quantifying stability.

Hence, our ML pipeline outperforms other approaches by jointly tackling three challenges of working with behavioral data – multiple modalities (*i.e.* collected from various data sources), high dimensionality with correlated features, and small sample sizes (resulting from logistical and privacy-related limitations during data collection).

Table 2. Comparing our method for detecting Post-semester Depression with 2 Baselines – K-Nearest Neighbors and Lasso. Our method performs better than the two baselines for all feature sets by employing a more stringent and robust feature selection strategy that consistently selects fewer but useful features. KEY – “N”: Sample Size, “F1”: F1 score of the "depression" class.

Feature Set	**N**	**Total Features**	**Method**	**Model Parameters**	**Accuracy**	**F1**	**No. of Features Selected (avg. across folds)**
Bluetooth	114	3202	KNN	K=2	53.5	.18	N/A
			Lasso	Alpha=0.7	60.5	.52	229
			Our Method	NRL (C=0.5, scaling=0.5, sample_fraction=0.80, selection_threshold=0.20) Model = GBC	**69.3**	**.64**	**73**
Calls	108	606	KNN	K=1	58.3	.46	N/A
			Lasso	Alpha=1.0	55.6	.33	57
			Our Method	NRL (C=0.5, scaling=0.7, sample_fraction=0.80, selection_threshold=0.80) Model = LogR (C=0.5 - same as NRL)	**68.5**	**.59**	**7**
Campus Map	110	23873	KNN	K=8	61.8	.46	N/A
			Lasso	Alpha=0.9	57.3	.53	140
			Our Method	NRL (C=0.3, scaling=0.5, sample_fraction=0.80, selection_threshold=0.20) Model = GBC	**68.2**	**.66**	**63**
Location	105	10238	KNN	K=7	61.9	.56	N/A
			Lasso	Alpha=0.3	50.5	.46	513
			Our Method	NRL (C=0.35, scaling=0.5, sample_fraction=0.85, selection_threshold=0.60) Model = LogR (C=0.35 - same as NRL)	**69.5**	**.62**	**10**
Phone Usage	111	15447	KNN	K=8	60.4	.35	N/A
			Lasso	Alpha=0.3	47.7	.37	261
			Our Method	NRL (C=0.6, scaling=0.5, sample_fraction=0.80, selection_threshold=0.50) Model = GBC	**70.3**	**.75**	**3**
Sleep	107	5890	KNN	K=10	49.5	.41	N/A
			Lasso	Alpha=1.0	44.9	.34	282
			Our Method	NRL (C=0.6, scaling=0.45, sample_fraction=0.80, selection_threshold=0.20) Model = GBC	**69.2**	**.66**	**74**
Steps	107	3055	KNN	K=9	**66.4**	.50	N/A
			Lasso	Alpha=0.3	62.6	.57	305
			Our Method	NRL (C=0.75, scaling=0.6, sample_fraction=0.80, selection_threshold=0.20) Model = LogR (C=0.75 - same as NRL)	63.6	**.53**	**80**
All	79	62311	KNN	K=3	64.6	.58	N/A
			Lasso	Alpha=0.7	59.5	.53	480
		Predictions from the 7 feature sets	Our Method	Combined predictions from the all seven 1-feature set models using AdaBoost (n_estimators=100)	**82.3**	**.78**	**310**

6.3 Implications for Longitudinal Studies and Opportunities to Improve Model Performance

Section 2 and table 1 show that our depression detection results are either better than or comparable to the current state-of-the-art. Further, our change of depression detection and depression prediction extend the current state-of-the-art. However, depression detection and prediction using mobile and wearable sensing is a fairly novel area of research, and there are significant opportunities to improve the accuracy of our models in future work. To this end, we have identified two possible kinds of sources of errors: (1) Errors that occur due to modeling, and (2) Errors that occur due to poor quantity or quality of data collected. Opportunities to mitigate these errors are described below.

The small sample size of our dataset contributes greatly to the errors that occur due to modeling. Increasing the sample size for training by collecting data from more people will increase the robustness and generalizability of our models and reduce error due to variance (*i.e.*, error due to small fluctuations in the training data), thereby improving accuracy. For this study, we started out with 188 participants but were left with 138 participants by the end of the study. 50 participants either dropped out, failed to answer depression questionnaires, or were missing much of their passively collected data due to technical issues. Hence, in order to increase the sample size, researchers will have to take a multi-pronged approach by (1) recruiting more participants, (2) encouraging compliance and reducing drop-out rates by offering additional or more engaging incentives (*e.g.*, interventions to improve their wellbeing), and (3) improving quantity or quality of data collected. Further, some participants may exhibit behavioral symptoms that are different from the rest of the population. Hence, in the future, researchers should investigate building personal models for each participant, such that each personal model contains weekly samples from 1 participant only, in order to predict the weekly depression labels for that participant. This kind of study will be challenging though, since self-report data will have to be collected over a much longer period of time.

We are currently repeating this study with a new cohort of first year undergraduate students from the same University and a subset of the now second year undergraduate students whose data was used in the analysis presented in this paper. This will allow us to compare behaviors from the same participants 1 year apart and their effect on depression, as well as build more stable models by training on a larger sample size and reporting test accuracies. This study is also being repeated at another University which will allow us to compare behavioral symptoms of depression and test the validity of our models across universities. These new studies will collect more frequent ground truth to allow us to improve and better understand our models for predicting depression in advance. We have also significantly improved our system and protocol for monitoring data collection daily throughout the study, which should greatly improve the quality of data collected.

To improve the quantity or quality of passively collected data, researchers need to monitor data collection daily and promptly address technical issues that cause noisy or missing data, as they arise. To this end, we have implemented a dashboard that shows us the amount of data received by the server from each participant daily. This allows us to reach out to participants and resolve data collection or data transmission issues that are causing missing data. In addition, efforts to encourage compliance and reduce drop-out rates will help improve the quantity or quality of data collected through questionnaires.

6.4 Implications for Interventions

Our machine learning approach enables building behavior models for early detection and prediction of change in depression without frequent ground truth data. This provides opportunities for timely interventions and treatments. We have discussed the implications of this work with mental health experts. They seem very excited about this research as they believe that this system can help them screen students for depression more efficiently. They also want to help us take this research forward by identifying modified behaviors that can be targeted during behavioral change interventions to improve depressive symptoms. In our sample of 138 participants, 40.6% students were found to have depressive symptoms post-semester, however only 17.4% students self-reported seeking counseling and psychological services. In subsequent studies with in-built interventions, we plan to use our system to identify students with depression and reach out to them through the student counseling center. Detecting post-semester depression allows us to identify students who may have a depressive disorder at the end of the semester. Detecting change in depression allows us to identify students who have worsened, such that we can intervene urgently and more aggressively if needed. Models for detecting change in depression may also be more sensitive to changes resulting from interventions, and hence, better at evaluating their effectiveness. Further, since our models are understandable, that is, they are built using meaningful behavioral features, they can be used to inform therapists treating students about the relationship between the students' behaviors and depression. As a result, therapists will be able to make more informed choices about which interventions would be most effective for each student. Students will also be able to participate in technological self-help interventions. For example, students can be shown visualizations of their sensed behaviors (features from our model) and their relationship with depression, thereby enabling guided self-reflection and planning for behavioral change.

The prediction models that predict post-semester depressive state and change in depression weekly, enable us to reach out to students who may be at-risk for depression as early as 1 to 5 weeks into a semester, in order to execute interventions to preempt depressive symptoms. However, we also find that the performance of these models trained at the end of each week varies over the 16 week period, instead of monotonically increasing. While this is expected behavior (and seen in previous work [11]) due to the weeks having semantic meaning, it makes it harder for university staff carrying out interventions to trust the output of the model at the end of any one week. Hence, to address this problem, we carried out additional analysis (*i.e.*, majority voting) and accordingly suggest an intervention strategy that utilizes our models. That is, we recommend that instead of trusting the output of one model at the end of a specific week, university staff should contact students predicted to be at-risk at the end of each week by a majority of *all* the models trained until that time point. We show that using this strategy would result in more stable accuracy and F1 values across the 16 weeks of the semester, and can thus be trusted more.

Detecting and monitoring depression in a large sample of students can also help inform policy changes at the university level, such as increasing outreach for psychological services, hiring more mental health professionals, and deciding drop deadlines for courses.

6.5 Implications for Privacy and Technical Limitations

The results of our feature ablation study show that we do *not* need data from all the sensor streams we recorded. In fact, combining features from fewer sensor streams often leads to better performance. For example, for detecting post-semester depression, a model containing features from all 7 sensors give us an accuracy of 82.3% while a model containing data from 4 sensors gives us an accuracy of 85.7%. This demonstrates an opportunity for algorithms that minimize data collection burden (*e.g.*, privacy, data transfer rate) while maximizing value (*i.e.*, model performance metrics like accuracy) for detecting mental health outcomes. As an example, consider our detection results. In both our detection outcomes, the best set model did not include Location, while for one outcome, it did include Campus Map. This means that from a privacy perspective and from a battery usage perspective, detailed and granular Location data is not needed, and instead human-understandable location (*i.e.*, Campus Map) labels are sufficient for well-performing models. As stated earlier, the human cost of obtaining Campus Map and Calls features is higher than for the other feature sets. Anyone implementing a detection or prediction system like the ones we have proposed in this paper, has to trade off this burden against the loss in accuracy that they might induce (*e.g.*, 3.2% loss in post-semester detection, and 8.9% loss in detecting change). Further, any features or feature sets that do not contribute to our best models means a reduction in the amount of data transferred from the phone to a back-end server. This also reduces battery usage, and potential financial costs to the participant depending on the data plan they have paid for.

To optimize for these types of burdens, Early *et al.* [28] present a method that dynamically chooses sensors and switches between them during data collection, thereby reducing data collection costs while achieving equivalent or better model performance. This method can be extended to our work in detecting mental health outcomes in college students.

6.6 Extending to Other Health Outcomes and Opportunities for Combining with Verbal and Non-Verbal Behaviors, and Genomic Data

We evaluated our ML pipeline in the context of depression, but it can be generalized to any chronic and longitudinal health problem. Further, depression has temperamental (cognitive), environmental (*e.g.*, childhood experiences, lifestyle), and genetic and physiological prognostic and risk factors [4, 7]. While we are able to detect depression by sensing daily behaviors, incorporating verbal and non-verbal behaviors and genomic data into our model will lead to a more holistic and unified model of depression [7]. This can help us predict depression before its onset more accurately, estimate prognosis after onset, and develop a better understanding of depression and its causes, thereby enabling more effective treatments and interventions for depression. We can do this by capturing cognitive (*e.g.,* negative beliefs [6]) and environmental (*e.g.*, abuse) factors using verbal behaviors from ecological momentary assessments [65] and social media posts [18], physiological (*e.g.*, response to stress) factors using wearable physiological sensors (*e.g.*, heart rate sensors) and hormonal testing (*e.g.*, saliva testing for stress hormones), and genetic factors using genomic sequencing. Large initiatives such as the UCLA Depression Grand Challenge[16] and the Precision Medicine Initiative[17] are already working on combining these different sources of data to detect and understand depression and other health-related outcomes. We plan to contribute to these initiatives by open sourcing our feature extraction library which will allow researchers to extract tens of thousands of behavioral and behavioral change features from a wide variety of sensor streams.

7 CONCLUSION

In this paper, we present a new feature selection approach that allows us to select meaningful features even when the number of features is significantly larger than the sample size. This approach enables models that detect depression at specific time points while considering a large set of features computed over the previous several weeks. We evaluate our approach by identifying students that have post-semester depressive symptoms using data collected over one semester (16 weeks) from the smartphones and fitness trackers of 138 college students, and achieve an accuracy of 85.7%. Further, we detect whether students' depressive symptom severity changed with an accuracy of 85.4%, and the levels of change with an accuracy of 82.9%.

Models that detect change in depression are novel, and will likely be better at evaluating interventions than diagnostic models. Finally, our work is the first to demonstrate that it is possible to predict depression several weeks in advance with an accuracy of >80% (*e.g.*, 81.3%, 11 weeks before the end of the semester). Hence, our work has significant implications for depression detection and monitoring, prediction before onset, and longitudinal symptom monitoring in-the-wild. Ultimately, it creates the potential for technology-mediated interventions that support the diagnosis, treatment, and prevention of depression. For example, a system built on data from these sensors can provide real-time feedback and alert the user before a depressive episode occurs. Such interventions could help increase awareness and motivate students to seek treatment and affect behavior change.

[16] https://grandchallenges.ucla.edu/depression/

[17] https://ghr.nlm.nih.gov/primer/precisionmedicine/initiative

In the future, features related to daily behaviors from our work can be combined with features related to verbal and non-verbal behaviors, and genomic data to develop a better understanding of depression and its causes, predict depression before its onset and prognosis after onset, thereby enabling more effective and personalized treatments and interventions for depression.

8 ACKNOWLEDGMENTS

This material is based upon work supported by Carnegie Mellon University's (CMU) Provost's Office, CMU's Center for Machine Learning and Health Fellowship, and the Carnegie Bosch Initiative.

REFERENCES

[1] Sharifa Alghowinem, Roland Goecke, Jeffrey F Cohn, Michael Wagner, Gordon Parker, and Michael Breakspear. 2015. Cross-cultural detection of depression from nonverbal behaviour. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 1. IEEE, 1–8.

[2] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parker, and Michael Breakspear. 2013. Eye movement analysis for depression detection. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 4220–4224.

[3] Sharifa Alghowinem, Roland Goecke, Michael Wagner, Gordon Parkerx, and Michael Breakspear. 2013. Head pose and movement analysis as an indicator of depression. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 283–288.

[4] American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

[5] Sangwon Bae, Anind K Dey, and Carissa A Low. 2016. Using passively collected sedentary behavior to predict hospital readmission. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 616–621.

[6] Aaron T Beck. 1979. *Cognitive therapy of depression*. Guilford press.

[7] Aaron T Beck and Keith Bredemeier. 2016. A unified model of depression: Integrating clinical, cognitive, biological, and evolutionary perspectives. *Clinical Psychological Science* 4, 4 (2016), 596–619.

[8] Aaron T Beck, Robert A Steer, and Gregory K Brown. 1996. Beck depression inventory-II. *San Antonio* 78, 2 (1996), 490–8.

[9] Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, and Andrew T Campbell. 2015. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric rehabilitation journal* 38, 3 (2015), 218.

[10] Robert Bryll, Ricardo Gutierrez-Osuna, and Francis Quek. 2003. Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets. *Pattern recognition* 36, 6 (2003), 1291–1302.

[11] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM, 1293–1304.

[12] Zhenyu Chen, Yiqiang Chen, Lisha Hu, Shuangquan Wang, Xinlong Jiang, Xiaojuan Ma, Nicholas D Lane, and Andrew T Campbell. 2014. ContextSense: unobtrusive discovery of incremental social context using dynamic bluetooth data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 23–26.

[13] Zhenyu Chen, Yiqiang Chen, Shuangquan Wang, Junfa Liu, Xingyu Gao, and Andrew T Campbell. 2013. Inferring social contextual behavior from bluetooth traces. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM, 267–270.

[14] Philip I Chow, Karl Fua, Yu Huang, Wesley Bonelli, Haoyi Xiong, Laura E Barnes, and Bethany A Teachman. 2017. Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students. *Journal of medical Internet research* 19, 3 (2017).

[15] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. 2009. Detecting depression from facial actions and vocal prosody. In *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 1–7.

[16] Jesse D Cook, Michael L Prairie, and David T Plante. 2017. Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: a comparison against polysomnography and wrist-worn actigraphy. *Journal of affective disorders* 217 (2017), 299–305.

[17] Ewa K Czyz, Adam G Horwitz, Daniel Eisenberg, Anne Kramer, and Cheryl A King. 2013. Self-reported barriers to professional help seeking among college students at elevated risk for suicide. *Journal of American College Health* 61, 7 (2013), 398–406.

[18] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 13 (2013), 1–10.

[19] Massimiliano de Zambotti, Aimee Goldstone, Stephanie Claudatos, Ian M Colrain, and Fiona C Baker. 2018. A validation study of Fitbit Charge 2âĎć compared with polysomnography in adults. *Chronobiology international* 35, 4 (2018), 465–476.

[20] Kadir Demirci, Mehmet Akgönül, and Abdullah Akpinar. 2015. Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students. *Journal of behavioral addictions* 4, 2 (2015), 85–92.

[21] Thomas G Dietterich. 1996. Statistical tests for comparing supervised classification learning algorithms. *Oregon State University Technical Report* 1 (1996), 1–24.

[22] Afsaneh Doryab. 2018. Identifying Symptoms Using Technology. In *Technology and Adolescent Mental Health*. Springer, 135–153.

[23] Afsaneh Doryab, Jun-Ki Min, Jason Wiese, John Zimmerman, and Jason I Hong. 2014. Detection of Behavior Change in People with Depression.. In *AAAI Workshop: Modern Artificial Intelligence for Health Analytics*.

[24] Afsaneh Doryab, Daniella K Villalba, Prerna Chikersal, Janine M Dutcher, Michael Tumminia, Xinwen Liu, Sheldon Cohen, Kasey Creswell, Jennifer Mankoff, John D Creswell, et al. 2019. Identifying Behavioral Phenotypes of Loneliness and Social Isolation with Passive Sensing: Statistical Analysis, Data Mining and Machine Learning of Smartphone and Fitbit Data. *JMIR mHealth and uHealth* 7, 7 (2019), e13209.

[25] David JA Dozois, Keith S Dobson, and Jamie L Ahnberg. 1998. A psychometric evaluation of the Beck Depression Inventory–II. *Psychological assessment* 10, 2 (1998), 83.

[26] Rakkrit Duangsoithong and Terry Windeatt. 2010. Bootstrap feature selection for ensemble classifiers. In *Industrial Conference on Data Mining*. Springer, 28–41.

[27] Nathan Eagle, Alex Sandy Pentland, and David Lazer. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences* 106, 36 (2009), 15274–15278.

[28] Kirstin Early, Stephen E Fienberg, and Jennifer Mankoff. 2016. Test time feature ordering with FOCUS: interactive predictions with minimal user burden. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 992–1003.

[29] Daniel Eisenberg, Ezra Golberstein, and Sarah E Gollust. 2007. Help-seeking and access to mental health care in a university student population. *Medical care* 45, 7 (2007), 594–601.

[30] Daniel Eisenberg, Ezra Golberstein, and Justin B Hunt. 2009. Mental health and academic success in college. *The BE Journal of Economic Analysis & Policy* 9, 1 (2009).

[31] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*, Vol. 96. 226–231.

[32] Asma Ahmad Farhan, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jin Lu, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis, and Bing Wang. 2016. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data.. In *Wireless Health*. 30–37.

[33] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. 2015. AWARE: mobile context instrumentation framework. *Frontiers in ICT* 2 (2015), 6.

[34] Darcy Gruttadaro and Dana Crudo. 2012. College students speak: A survey report on mental health. *Retrieved from National Alliance on Mental Illness website: https://www. nami. org/About-NAMI/Publications-Reports/Survey-Reports/College-Students-Speak_A-Survey-Report-on-Mental-H. pdf* (2012).

[35] Eric Heiligenstein, Greta Guenther, Ken Hsu, and Kris Herman. 1996. Depression and academic impairment in college students. *Journal of American College Health* 45, 2 (1996), 59–64.

[36] Tim Bodyka Heng, Ankit Gupta, and Chris Shaw. 2018. FitViz-Ad: A Non-Intrusive Reminder to Encourage Non-Sedentary Behaviour. *Electronic Imaging* 2018, 1 (2018), 332–1.

[37] Alketa Hysenbegasi, Steven L Hass, and Clayton R Rowland. 2005. The impact of depression on the academic productivity of university students. *Journal of Mental Health Policy and Economics* 8, 3 (2005), 145.

[38] Jyoti Joshi, Roland Goecke, Gordon Parker, and Michael Breakspear. 2013. Can body expressions contribute to automatic depression analysis?. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 1–7.

[39] Raghavendra Katikalapudi, Sriram Chellappan, Frances Montgomery, Donald Wunsch, and Karl Lutzen. 2012. Associating internet usage with depressive behavior among college students. *IEEE Technology and Society Magazine* 31, 4 (2012), 73–80.

[40] Ronald C Kessler, Patricia Berglund, Guilherme Borges, Matthew Nock, and Philip S Wang. 2005. Trends in suicide ideation, plans, gestures, and attempts in the United States, 1990-1992 to 2001-2003. *Jama* 293, 20 (2005), 2487–2495.

[41] Jeremy Kisch, E Victor Leino, and Morton M Silverman. 2005. Aspects of suicidal behavior, depression, and treatment in college students: Results from the Spring 2000 National College Health Assessment Survey. *Suicide and Life-Threatening Behavior* 35, 1 (2005), 3–13.

[42] Kurt Kroenke and Robert L Spitzer. 2002. The PHQ-9: a new depression diagnostic and severity measure. *Psychiatric annals* 32, 9 (2002), 509–515.

[43] Min Kwon, Joon-Yeop Lee, Wang-Youn Won, Jae-Woo Park, Jung-Ah Min, Changtae Hahn, Xinyu Gu, Ji-Hye Choi, and Dai-Jin Kim. 2013. Development and validation of a smartphone addiction scale (SAS). *PloS one* 8, 2 (2013), e56936.

[44] Janna Mantua, Nickolas Gravel, and Rebecca Spencer. 2016. Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography. *Sensors* 16, 5 (2016), 646.

[45] Abhinav Mehrotra and Mirco Musolesi. 2018. Using Autoencoders to Automatically Extract Mobility Features for Predicting Depressive States. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 127.

[46] Nicolai Meinshausen and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, 4 (2010), 417–473.

[47] Krista Merry and Pete Bettinger. 2019. Smartphone GPS accuracy study in an urban environment. *PloS one* 14, 7 (2019), e0219890.

[48] Jan Mielniczuk and Paweł Teisseyre. 2014. Using random subspace method for prediction and variable importance assessment in linear regression. *Computational Statistics & Data Analysis* 71 (2014), 725–742.

[49] Scott M Monroe, George M Slavich, Leandro D Torres, and Ian H Gotlib. 2007. Major life events and major chronic difficulties are differentially associated with history of major depressive episodes. *Journal of abnormal psychology* 116, 1 (2007), 116.

[50] Stuart A Montgomery and MARIE Åsberg. 1979. A new depression scale designed to be sensitive to change. *The British journal of psychiatry* 134, 4 (1979), 382–389.

[51] Tom Nicolai and Holger Kenn. 2006. Towards detecting social situations with Bluetooth. In *Adjunct Proceedings Ubicomp*.

[52] Matthew K Nock and Ronald C Kessler. 2006. Prevalence of and risk factors for suicide attempts versus suicide gestures: analysis of the National Comorbidity Survey. *Journal of abnormal psychology* 115, 3 (2006), 616.

[53] David Nutt, Sue Wilson, and Louise Paterson. 2008. Sleep disorders as core symptoms of depression. *Dialogues in clinical neuroscience* 10, 3 (2008), 329.

[54] Yongjun Piao, Minghao Piao, Cheng Hao Jin, Ho Sun Shon, Ji-Moon Chung, Buhyun Hwang, and Keun Ho Ryu. 2015. A new ensemble method with feature space partitioning for high-dimensional data classification. *Mathematical Problems in Engineering* 2015 (2015).

[55] William H Press and George B Rybicki. 1989. Fast algorithm for spectral analysis of unevenly sampled data. *The Astrophysical Journal* 338 (1989), 277–280.

[56] I Qualtrics. 2013. Qualtrics. *Provo, UT, USA* (2013).

[57] Jane M Rondina, Tim Hahn, Leticia de Oliveira, Andre F Marquand, Thomas Dresler, Thomas Leitner, Andreas J Fallgatter, John Shawe-Taylor, and Janaina Mourao-Miranda. 2013. SCoRSâĂŤA method based on stability for feature selection and mapping in neuroimaging. *IEEE transactions on medical imaging* 33, 1 (2013), 85–98.

[58] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition & Emotion* 18, 8 (2004), 1121–1133.

[59] Sohrab Saeb, Emily G Lattie, Stephen M Schueller, Konrad P Kording, and David C Mohr. 2016. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ* 4 (2016), e2537.

[60] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015).

[61] Sara E Schaefer, Cynthia Carter Ching, Heather Breen, and J Bruce German. 2016. Wearing, thinking, and moving: testing the feasibility of fitness tracking with urban youth. *American Journal of Health Education* 47, 1 (2016), 8–16.

[62] Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency. 2013. Investigating voice quality as a speaker-independent indicator of depression and PTSD.. In *Interspeech*. 847–851.

[63] Stefan Scherer, Giota Stratou, and Louis-Philippe Morency. 2013. Audiovisual behavior descriptors for depression assessment. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 135–140.

[64] Mohammed Senoussaoui, Milton Sarria-Paja, João F Santos, and Tiago H Falk. 2014. Model fusion for multimodal depression classification and level detection. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 57–63.

[65] Saul Shiffman, Arthur A Stone, and Michael R Hufford. 2008. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.* 4 (2008), 1–32.

[66] Grace Shin, Yuanyuan Feng, Mohammad Hossein Jarrahi, and Nicci Gafinowitz. 2018. Beyond novelty effect: a mixed-methods exploration into the motivation for long-term activity tracker use. *JAMIA Open* 2, 1 (2018), 62–72.

[67] Karen L Smarr and Autumn L Keefer. 2011. Measures of depression and depressive symptoms: Beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS),

Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9). *Arthritis care & research* 63, S11 (2011).

[68] Giota Stratou, Stefan Scherer, Jonathan Gratch, and Louis-Philippe Morency. 2013. Automatic nonverbal behavior indicators of depression and PTSD: Exploring gender differences. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on.* IEEE, 147–152.

[69] Ashleigh Sushames, Andrew Edwards, Fintan Thompson, Robyn McDermott, and Klaus Gebel. 2016. Validity and reliability of Fitbit Flex for step count, moderate to vigorous physical activity and activity energy expenditure. *PloS one* 11, 9 (2016), e0161224.

[70] Tan-Hsu Tan, Munkhjargal Gochoo, Ke-Hao Chen, Fu-Rong Jean, Yung-Fu Chen, Fu-Jin Shih, and Chiung Fang Ho. 2014. Indoor activity monitoring system for elderly using RFID and Fitbit Flex wristband. In *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI).* IEEE, 41–44.

[71] John D Teasdale. 1997. The relationship between cognition and emotion: The mind-in-place in mood disorders. (1997).

[72] A Trajman and RR Luiz. 2008. McNemar $\chi 2$ test revisited: comparing sensitivity and specificity of diagnostic examinations. *Scandinavian journal of clinical and laboratory investigation* 68, 1 (2008), 77–80.

[73] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, and Steffi Weidt. 2016. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR mHealth and uHealth* 4, 3 (2016).

[74] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* ACM, 3–14.

[75] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T Campbell. 2015. SmartGPA: how smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing.* ACM, 295–306.

[76] Rui Wang, Weichen Wang, Min SH Aung, Dror Ben-Zeev, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A Scherer, et al. 2017. Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 110.

[77] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 43.

[78] Yilun Wang, Zhiqiang Li, Yifeng Wang, Xiaona Wang, Junjie Zheng, Xujuan Duan, and Huafu Chen. 2015. A novel approach for stable selection of informative redundant features from high dimensional fMRI data. *arXiv preprint arXiv:1506.08301* (2015).

[79] Zhixian Yan, Jun Yang, and Emmanuel Munguia Tapia. 2013. Smartphone bluetooth based social sensing. In *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication.* ACM, 95–98.

[80] Jie Zhang, Zhigen Zhao, Kai Zhang, and Zhi Wei. 2017. A feature sampling strategy for analysis of high dimensional genomic data. *IEEE/ACM transactions on computational biology and bioinformatics* 16, 2 (2017), 434–441.

A APPENDIX: QUANTIFYING MISSING FEATURES

Missing features were handled using 2 steps:

(1) **Exclusion:** All features that were missing for more than 30 participants are excluded. Further, if a participant was missing more than 20% of all features from a feature set, we removed that participant.
(2) **Imputation:** The remaining feature matrix still contained some missing cells (*i.e.* some features were missing for certain participants), which were imputed as '-1'.

Table 3 shows the size of the feature matrix (*i.e.* number of participants * number of features) before and after the exclusion step of missing features handling. Missing features in the resulting feature matrix (of dimensions $N*NFeats\_input$) will be imputed as '-1' and this feature matrix will be given as input to feature selection and modeling. These missing features are not necessarily missing at random, and the '-1' may be semantically meaningful (see section 4.2).

Figure 6 shows the percentage of imputations *i.e.* the percentage of cells imputed in the feature matrix containing 'All' features, or the feature matrices filtered by features from each week or each epoch. There is one figure for each feature set. For Bluetooth, Campus Map, Location, and Phone Usage, < 5% of the feature matrix was imputed due to missing features. Whereas, for Calls, Sleep, and Steps, 5-10% was often imputed due to missing features. There are more missing features for Calls because we cannot differentiate between no calls or the calls sensor not working, and missing calls could simply mean that calls were made in that time period. Further, Sleep and Steps have increasingly more missing features as the semester progresses, probably because participants find it harder to wear the Fitbit as the semester progresses due to increased workload and wearing off of the novelty effect [66]. Sleep features were also missing more often from the afternoons and evenings. No other distinct patterns were seen across the weeks/ epochs.

Table 3. The extracted (raw) feature matrix had dimensions $N\_raw*NFeats\_raw$. Then, all features that were missing for more than 30 participants and all participants that had more than 20% missing features were excluded. The resulting feature matrix had dimensions $N*NFeats\_input$.

	Before Missing Data Handling		**After Missing Data Handling But Before Feature Selection**	
Feature Set	**No. of Participants (N_raw)**	**No. of Features (NFeats_raw)**	**No. of Participants (N)**	**No. of Features (NFeats_input)**
Bluetooth	138	4275	114	3202
Calls	138	2850	108	606
Campus map	138	26790	110	23873
Location	138	10830	105	10238
Phone Usage	138	16815	111	15447
Sleep	138	12255	107	5737
Steps	138	3705	107	3055

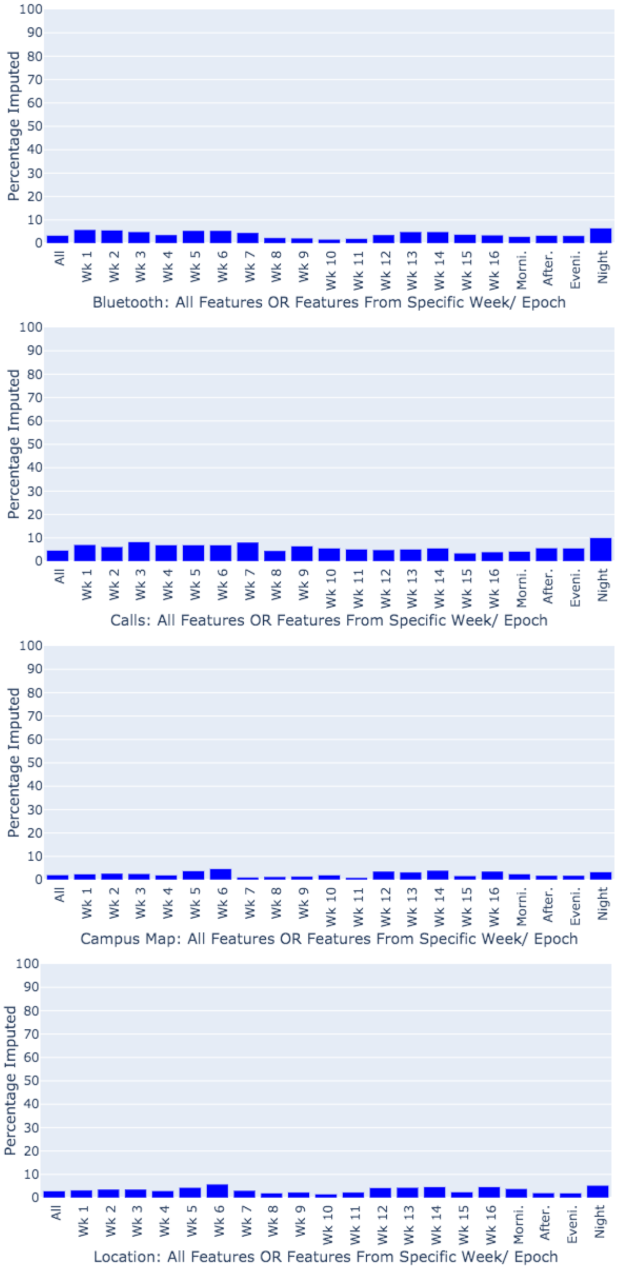


Fig. 6. *(contd.)* Percentage of missing features ('cells') in the feature matrix given as input to feature selection and modeling. We filter the feature matrix by week number and epoch (morning, afternoon, evening, and night) to see if there is any pattern to the 'missingness'.

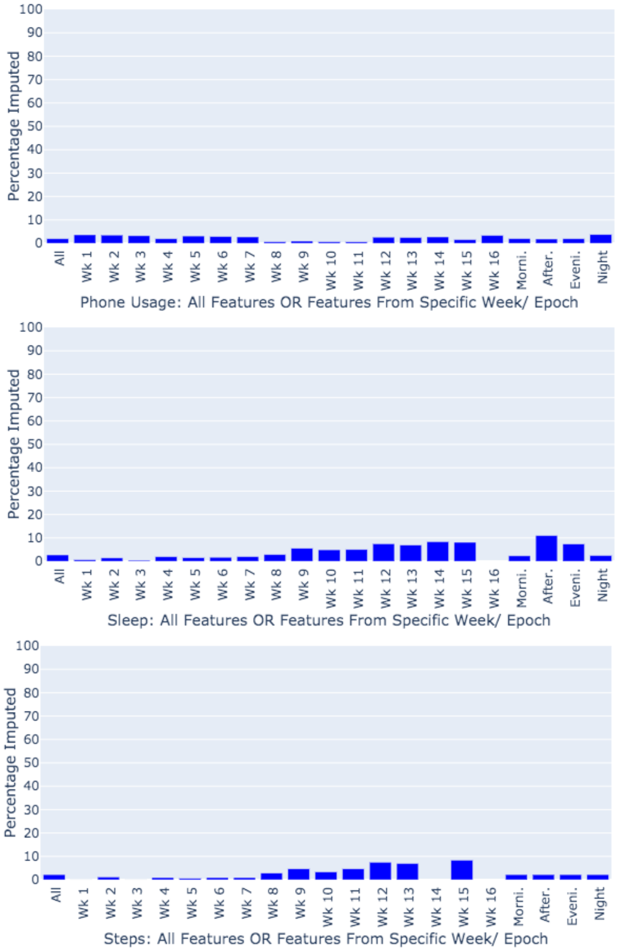


Fig. 6. *(contd.)* Percentage of missing features ('cells') in the feature matrix given as input to feature selection and modeling. We filter the feature matrix by week number and epoch (morning, afternoon, evening, and night) to see if there is any pattern to the 'missingness'.

B APPENDIX: ADDITIONAL EVALUATION METRICS FOR OUR MODELS

Tables 4 and 5 presents additional evaluation metrics for each 1-feature set detection model, the "all" and "best" combined detection models, and the earliest of the top-5 combined prediction models for each of the two outcomes.

Tables 6 to 11 present the confusion matrices for the "all" and "best" combined detection models, and the earliest of the top-5 combined prediction models for each of the two outcomes. The confusion matrix contains the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN).

Note – N: Number of samples, NS: number of feature-sets combined, NFeats: Avg. number of feature selected across all folds from total number of features in feature-set, MCC: Matthews Correlation Coefficient, F1: F1-score, P: Precision, R: Recall.

Table 4. Additional evaluation metrics for **Post-semester Depression**.

Task and Model	N	NS/ NFeats	Accuracy	F1 Dep.	P Dep.	R Depression	F1 No Dep.	P No Dep.	R No Dep.
Detection Using Bluetooth	114	73 out of 3202 features sel.	69.3	0.64	0.64	0.6	0.72	0.72	0.76
Detection Using Calls	108	7 out of 606 features sel.	68.5	0.59	0.59	0.78	0.8	0.8	0.62
Detection Using Campus Map	110	63 out of 23873 features sel.	68.2	0.66	0.66	0.56	0.7	0.7	0.77
Detection Using Location	105	10 out of 10238 features sel.	69.5	0.62	0.62	0.8	0.8	0.8	0.61
Detection Using Phone Usage	111	3 out of 15447 features sel.	70.3	0.75	0.75	0.45	0.69	0.69	0.89
Detection Using Sleep	107	74 out of 5890 features sel.	69.2	0.66	0.66	0.49	0.71	0.71	0.83
Detection Using Steps	107	80 out of 3055 features sel.	63.6	0.53	0.53	0.55	0.7	0.7	0.69
Detection Using All Feature-Sets	79	7 feature-sets	82.3	0.78	0.78	0.78	0.85	0.85	0.85
Detection Using Best Feature-Sets	84	4 feature-sets	85.7	0.82	0.82	0.82	0.88	0.88	0.88
Prediction by the Earliest of Top-5	80	7 feature-sets	81.3	0.75	0.76	0.73	0.85	0.84	0.86

Table 5. Additional evaluation metrics for **Change in Depression**.

Task and Model	**N**	**NS/ NFeats**	**Accuracy**	**F1 Dep.**	**P Dep.**	**R Depression**	**F1 No Dep.**	**P No Dep.**	**R No Dep.**
Detection Using Bluetooth	114	83 out of 3202 features sel.	65.8	0.55	0.55	0.27	0.68	0.68	0.88
Detection Using Calls	108	16 out of 606 features sel.	64.8	0.5	0.5	0.39	0.71	0.71	0.79
Detection Using Campus Map	110	14 out of 23873 features sel.	79.1	0.8	0.8	0.59	0.79	0.79	0.91
Detection Using Location	105	25 out of 10238 features sel.	74.3	0.73	0.73	0.49	0.75	0.75	0.89
Detection Using Phone Usage	111	6 out of 15448 features sel.	73.9	0.68	0.68	0.56	0.77	0.77	0.84
Detection Using Sleep	107	6 out of 5890 features sel.	73.8	0.66	0.66	0.63	0.78	0.78	0.81
Detection Using Steps	107	34 out of 3055 features sel.	68.2	0.56	0.56	0.62	0.77	0.77	0.72
Detection Using All Feature-Sets	79	7 feature-sets	75.9	0.67	0.68	0.66	0.81	0.8	0.82
Detection Using Best Feature-Sets	82	4 feature-sets	85.4	0.8	0.83	0.77	0.88	0.87	0.9
Prediction by the Earliest of Top-5	84	7 feature-sets	88.1	0.81	0.92	0.73	0.91	0.87	0.96

Table 6. Confusion Matrix for the Model Containing All Sensors for Detecting Post-Semester Depression

		Model Output Label	
		No Depression	**Depression**
True Label	**No Depression**	40 (TN)	7 (FP)
	Depression	7 (FN)	25 (TP)

Table 7. Confusion Matrix for the Best Set Model for Detecting Post-Semester Depression

		Model Output Label	
		No Depression	**Depression**
True Label	**No Depression**	45 (TN)	6 (FP)
	Depression	6 (FN)	27 (TP)

Table 8. Confusion Matrix for the Earliest of Top-5 Models for Predicting Post-Semester Depression

		Model Output Label	
		No Depression	**Depression**
True Label	**No Depression**	43 (TN)	7 (FP)
	Depression	8 (FN)	22 (TP)

Table 9. Confusion Matrix for the Model Containing All Sensors for Detecting Change in Depression

		Model Output Label	
		No Depression	**Depression**
True Label	**No Depression**	41 (TN)	9 (FP)
	Depression	10 (FN)	19 (TP)

Table 10. Confusion Matrix for the Best Set Model for Detecting Change in Depression

		Model Output Label	
		Did not worsen	**Worsens**
True Label	**Did not worsen**	46 (TN)	5 (FP)
	Worsens	7 (FN)	24 (TP)

Table 11. Confusion Matrix for the Earliest of Top-5 Models for Predicting Change in Depression

		Model Output Label	
		Did not worsen	**Worsens**
True Label	**Did not worsen**	52 (TN)	2 (FP)
	Worsens	8 (FN)	22 (TP)

C APPENDIX: FEATURES SELECTED BY OUR MODELS

Table 12 lists the features selected in all folds of the leave-one-out feature selection when building 1-feature set models for detecting post-semester depression and change in depression. These features may help inform future research.

Table 12. Selected features from each 1-feature set model that were given as input to the Association Rule Mining (Apriori) algorithm.

Feature Set	Outcome	Features Selected in All Folds
Bluetooth	Post-semester	'Average number of scans of all devices of others in the afternoons in week 7', 'Average number of scans of all devices of others in the afternoons on weekends in week 1', 'Average number of scans of all devices of others in the afternoons on weekends in week 9', 'Average number of scans of all devices of others in the evenings in week 11', 'Average number of scans of all devices of others in the evenings on weekdays in week 9', 'Number of scans of least frequent device in the afternoons on weekends in week 1', 'Number of scans of least frequent device in the mornings on weekdays in week 9', 'Number of scans of least frequent device in the mornings on weekends in weeks 1-6', 'Number of scans of least frequent device in the nights on weekdays in week 14', 'Number of scans of least frequent device of others in the afternoons on weekends in week 1', 'Number of scans of least frequent device of others in the evenings on weekdays in week 3', 'Number of scans of least frequent device of others in the mornings in week 15', 'Number of scans of least frequent device of others in the nights in week 4', 'Number of scans of least frequent device of others in the nights on weekdays in week 14', 'Number of scans of least frequent device of others in the nights on weekdays in week 5', 'Number of scans of least frequent device of others in the nights on weekends in weeks 1-6', 'Number of scans of least frequent device of others on weekdays in week 3', 'Number of scans of most frequent device of others in week 6', 'Number of unique devices of self in the evenings on weekends in week 2', 'Number of unique devices of self in the evenings on weekends in week 5'

Table continued on the next page . . .

Table 12 continued from previous page

Feature Set	**Outcome**	**Features Selected in All Folds**
Bluetooth	Change	'Average number of scans of all devices of others in the afternoons in week 7', 'Average number of scans of all devices of others in the afternoons on weekends in week 1', 'Average number of scans of all devices of others in the afternoons on weekends in weeks 1-6', 'Average number of scans of all devices of others in the evenings in week 11', 'Average number of scans of all devices of others in the evenings on weekdays in week 9', 'Average number of scans of all devices of others in the evenings on weekends in week 4', 'Average number of scans of all devices of others in the mornings in week 1', 'Average number of scans of all devices of others in week 7', 'Number of scans of least frequent device in the afternoons on weekends in week 1', 'Number of scans of least frequent device in the mornings in week 3', 'Number of scans of least frequent device in the mornings on weekdays in week 9', 'Number of scans of least frequent device in the nights in week 15', 'Number of scans of least frequent device of others in the afternoons on weekends in week 1', 'Number of scans of least frequent device of others in the nights on weekdays in week 14', 'Number of scans of least frequent device of others in the nights on weekends in weeks 1-6', 'Number of scans of least frequent device of others on weekends in week 8', 'Number of scans of least frequent device on weekends in week 2', 'Number of unique devices of self in the afternoons on weekends in week 2', 'Number of unique devices of self in the evenings in week 6', 'Number of unique devices of self in the evenings on weekends in week 2', 'Number of unique devices of self in the evenings on weekends in week 5', 'Number of unique devices of self on weekends in week 8', 'Std number of scans of all devices of others in the afternoons on weekdays in week 1'
Calls	Post-semester	'Number of incoming calls in the evenings in week 13', 'Number of missed calls in the evenings in week 1'
Calls	Change	'Number of incoming calls in the evenings in week 11', 'Number of incoming calls in the evenings in week 13', 'Number of incoming calls in the evenings on weekdays in week 5'

Table continued on the next page . . .

Table 12 continued from previous page

Feature Set	**Outcome**	**Features Selected in All Folds**
Campus Map	Post-semester	'Maximum bout in residential apartments in the evenings in week 12', 'Maximum bout on-campus in week 11', 'Minimum bout off-campus in the mornings on weekdays in week 14', 'Minimum bout in green spaces in the evenings on weekdays in week 12', 'Minimum bout in green spaces in the mornings on weekdays in week 10', 'Minimum bout on-campus in the evenings on weekends in week 14', 'Minutes in athletic faciltiies in the afternoons on weekdays in week 2', 'Minutes in green spaces in the evenings on weekdays in week 6', 'Minutes in residential apartments in the evenings in week 12', 'Number of bouts 20min or more in residential apartments in the afternoons on weekends in week 10', 'Number of bouts 30min or more in residential halls in the afternoons on weekdays in week 14', 'Percent time in residential halls in the nights on weekends in week 8', 'Std bout in residential halls in the afternoons in week 9', 'Std bout off-campus in week 11'
Campus Map	Change	'Maximum bout in green spaces on weekdays in week 13', 'Std bout in residential halls in the afternoons in week 9'
Location	Post-semester	'Moving time percent in the nights on weekends in week 8'
Location	Change	'Home stay time percent 100m in the evenings in week 9', 'Home stay time percent 100m in the evenings on weekends in week 9', 'Mean length stay at significant locations in minutes (local clusters) in the afternoons on weekends in week 12', 'Number of of significant locations in the afternoons on weekends in week 12'
Phone Usage	Post-semester	'Number of times last "unlock" at hour 3 in the mornings on weekdays in week 9'
Phone Usage	Change	'Number of times last "unlock" at hour 3 in the mornings on weekdays in week 9'

Table continued on the next page . . .

Table 12 continued from previous page

Feature Set	**Outcome**	**Features Selected in All Folds**
Sleep	Post-semester	'Average length bout awake in week 1', 'End time maximum bout asleep in the nights on weekdays in week 2', 'End time maximum bout asleep on weekdays in week 11', 'End time minimum bout restless in the nights in week 12', 'End time minimum bout totalsleep in the nights in week 3', 'Maximum length bout awake in the mornings on weekdays in week 5', 'Maximum length bout awake in the nights on weekends in week 10', 'Minimum length bout awake in the nights on weekends in week 3', 'Number of asleep bouts in the mornings on weekdays in week 8', 'Number of restless bouts in the mornings on weekends in week 4', 'Number of unknown instances on weekends in week 12', 'Start time maximum bout asleep in the mornings in week 3', 'Start time maximum bout asleep in the mornings in weeks 1-6', 'Start time maximum bout restless in week 5', 'Start time maximum bout restless on weekdays in week 5', 'Start time minimum bout awake on weekdays in week 7', 'Start time minimum bout restless in the nights in week 12', 'Weak sleep efficiency in the nights in week 1'
Sleep	Change	'Number of awake on weekdays in week 10', 'Start time maximum bout asleep in the mornings in week 3', 'Sum length bout awake on weekdays in week 10'
Steps	Post-semester	'Average length active bout minutes in the afternoons on weekdays in week 6', 'Average length active bout minutes in the evenings on weekends in week 4', 'Average length active bout minutes in the mornings on weekends in week 3', 'Average length active bout minutes in the nights on weekdays in week 1', 'Maximum length active bout minutes in the afternoons in week 1', 'Maximum length active bout minutes in the evenings on weekends in weeks 1-16', 'Maximum length sedentary bout minutes in the afternoons in week 6', 'Maximum length sedentary bout minutes in the evenings on weekdays in week 12', 'Maximum length sedentary bout minutes in the nights on weekdays in week 3', 'Maximum step active bout in the evenings on weekdays in week 2', 'Maximum steps in the afternoons in week 3', 'Maximum steps in the nights in week 3', 'Maximum steps in the nights on weekdays in week 2', 'Number of active bout in the afternoons on weekdays in week 1', 'Number of active bout in the afternoons on weekends in week 3', 'Number of sedentary bout in the afternoons on weekdays in week 1', 'Number of sedentary bout in the nights in week 11'

Table continued on the next page ...

Table 12 continued from previous page

Feature Set	Outcome	Features Selected in All Folds
Steps	Change	'Maximum length sedentary bout minutes in the evenings on weekdays in week 12', 'Maximum steps in the afternoons in week 3', 'Minimum length sedentary bout minutes in the mornings on weekends in week 10', 'Number of active bout in the afternoons on weekends in week 3', 'Number of sedentary bout in the nights on weekends in week 11'

D APPENDIX: FEATURE ABLATION STUDY

Measuring the salience of each feature set can inform future research and enable depression detection models that are optimized for privacy, and technical limitations such as battery life and data transfer rate. While figure 4 indicates how salient or "useful" each feature set is on its own, it does not allow us to analyze the salience of a feature set when it is combined with other feature sets. It is important to analyze the latter because a feature set may not be significantly better on its own but can have a significant effect on accuracy when present in combination with other feature sets. For this purpose, we carried out a feature ablation study (see section 4.3.3). We tried 120 different combinations of feature sets and obtained their accuracies. Analyzing these results was not trivial. For example, we found no pattern in the accuracies of models obtained by removing 1 feature set at a time (*i.e.*, 6-feature sets models). Hence, for each feature set, we calculate the average accuracy of all models containing it, and report our findings below.

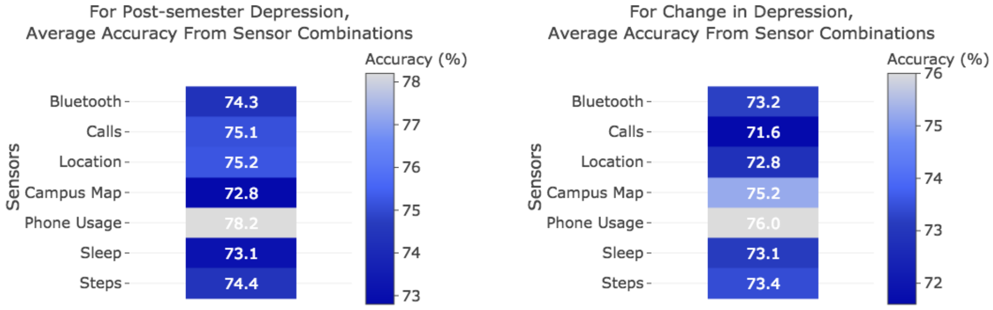
Figure 7a shows the average accuracy per feature set for detecting post-semester depression. "Phone Usage" (78.2%) has the highest average accuracy, and is closely followed by "Location" (75.2%) and "Calls" (75.1%). The feature ablation study also gave us a best set accuracy of 85.7% (N=84) using a model containing 4 feature sets: "Bluetooth", "Calls", "Phone Usage", "Steps".

Figure 7b shows the average accuracy per feature set for detecting change in depression. "Phone Usage" (76.0%) has the highest average accuracy, followed by "Campus Map" (75.2%). Also, a best set accuracy of 85.4% (N = 82) was obtained using a model containing 4 feature sets: "Bluetooth", "Campus Map", "Phone Usage", and "Sleep.

Our findings show that "Bluetooth", "Phone Usage" and "Location" or "Campus Map" (which is calculated using location) are salient across both outcomes. Previous work has repeatedly focused on the use of "Location" and "Phone Usage" as the most important sensors [59, 60], while "Bluetooth" has mostly been ignored. Our results indicate that "Bluetooth" may also convey interesting information and hence, should not be ignored.

Further, the "Campus Map" and "Calls" feature sets are harder to acquire than the other feature sets, since they require the researchers to input a map of the campus and participants to provide the phone numbers of their family members and friends. Hence, we are reporting how well our approach works without these two feature sets, below. This may help researchers decide if they want to exclude these two feature sets in the future.

The best accuracy obtained for detecting post-semester depression without these two feature sets is 82.5% (N = 80), and is obtained using "Bluetooth", "Location", "Phone Usage", "Sleep", and "Steps". This is slightly lower than the best overall accuracy for detecting post-semester depression which was 85.7% (N = 84). The best accuracy obtained for detecting change in depression without these two feature sets is 76.5% (N = 81), and is obtained using "Bluetooth", "Location", "Phone Usage", and "Steps". This is significantly lower than the best overall accuracy for detecting change in depression which was 85.4% (N = 82).



(a) Heat map indicating usefulness of sensors for detecting Post-semester Depression.

(b) Heat map indicating usefulness of sensors for detecting Change in Depression.

Fig. 7. Heat maps indicating the salience of feature sets derived from different sensors for detecting (a) Post-semester Depression, and (b) Change in Depression. For each feature set, we calculate the average accuracy of all combinations of feature sets containing that feature set.

E MODEL PARAMETER TUNING

To tune model parameters, we did a semi-greedy grid search. For each of the 7 feature sets, we tried Logistic Regression (LogR) as well as Gradient Boosting Classifier (GBC). We tune the parameters for these two models as follows:

(1) Take selection_threshold = 0.3, sample_fraction = 0.80, and scaling = 0.5. Vary C = [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0] and choose C = best_C_from_step_1 based on CV accuracy.
(2) Take C = best_C_from_step_1, sample_fraction = 0.80, and scaling = 0.5. Vary selection_threshold = [0.1, 0.2, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0] and choose best_selection_threshold_from_step_2 based on CV accuracy.
(3) Take C = best_C_from_step_1, selection_threshold = best_selection_threshold_from_step_2, and sample_fraction = 0.80. Vary scaling = [0.3, 0.4, 0.6, 0.7] and choose best_scaling_from_step_3 based on CV accuracy.
(4) Take C = best_C_from_step_1, selection_threshold = best_selection_threshold_from_step_2, and scaling = best_scaling_from_step_3. Vary sample_fraction = [0.75, 0.85] and choose best_sample_fraction_from_step_4 based on CV accuracy.
(5) Take selection_threshold = best_selection_threshold_from_step_2, scaling = best_scaling_from_step_3, and sample_fraction = best_sample_fraction_from_step_4. Vary C = [(best_C_from_step_1 + 0.05), (best_C_from_step_1 - 0.05)] and choose best_C_final based on CV accuracy.
(6) Take C = best_C_final, scaling = best_scaling_from_step_3, and sample_fraction = best_sample_fraction_from_step_4. Vary selection_threshold = [(best_selection_threshold_from_step_2 + 0.05), (best_selection_threshold_from_step_2 - 0.05)] and choose best_selection_threshold based on CV accuracy.
(7) Take C = best_C_final, selection_threshold = best_selection_threshold, and sample_fraction = best_sample_fraction_from_step_4. Vary scaling = [(best_scaling_from_step_3 + 0.05), (best_scaling_from_step_3 - 0.05)] and choose best_scaling based on CV accuracy.
(8) Take C = best_C_final, selection_threshold = best_selection_threshold, and scaling = best_scaling. Vary sample_fraction = [(best_sample_fraction_from_step_4 + 0.05),

(best_sample_fraction_from_step_4 - 0.05)] and choose best_sample_fraction based on CV accuracy.

(9) Final parameters are: C = best_C_final, selection_threshold = best_selection_threshold, scaling = best_scaling, and sample_fraction = best_sample_fraction.

STATEMENT OF PREVIOUS RESEARCH

The authors on this paper have also co-authored a paper titled “Identifying Behavioral Phenotypes of Loneliness and Social Isolation with Passive Sensing: A Three-fold Analysis” [24]. While there is a relationship between loneliness and depression, these are two different cognitive constructs and often have different dynamics. For example, in our sample of college students, depression rate almost triples from the beginning to the end of the semester while loneliness rate stays the same. This paper uses the same dataset described in section 3 and features described in section 4.1, to primarily do the following:

- Perform statistical analysis to understand the relationship between these features and loneliness in college students.
- Uses the Apriori algorithm to extract combined behavior patterns associated with loneliness.

Further, this paper uses the same pipeline described in section 4.3 to detect post-semester loneliness and change in loneliness. It does not consider any depression related outcomes, and does not attempt early prediction of post-semester loneliness. The paper on loneliness will cite this paper.